Published in final edited form as:

Med Decis Making. 2013 August ; 33(6): . doi:10.1177/0272989X13487605.

Patient and Societal Value Functions for the Testing Morbidities Index

John Shannon Swan, M.D.^{1,2}, Chung Yin Kong, Ph.D.^{1,2}, Janie M. Lee, M.D., M.Sc.^{1,2}, Omosalewa Akinyemi, M.D., M.P.H.^{1,3}, Elkan F. Halpern, Ph.D.^{1,2}, Pablo Lee, B.S.¹, Sergey Vavinskiy, Ph.D.^{4,5}, Olubunmi Williams, M.D., M.P.H.^{1,6}, Emilie S. Zoltick, B.A.^{1,7}, and Karen Donelan, Ed.M., Sc.D.^{1,2,8}

¹Massachusetts General Hospital Institute for Technology Assessment, Boston, MA

²Harvard Medical School, Boston, MA

³Department of Obstetrics and Gynecology, Wayne State University School of Medicine, Detroit, MI

⁴Indiana University Department of Radiology, Indianapolis, IN

⁵Indiana State Government, Indianapolis, IN

⁶Renal Division, Emory University, Atlanta, GA

⁷Boston University School of Public Health, Boston, MA

⁸Mongan Institute for Health Policy and Massachusetts General Hospital, Boston, MA

Abstract

Background—We developed preference-based and summated scale scoring for the Testing Morbidities Index (TMI) classification, which addresses short-term effects on quality of life from diagnostic testing before, during and after a testing procedure.

Methods—The two TMI value functions utilize multiattribute value techniques; one is patient-based and the other has a societal perspective. 206 breast biopsy patients and 466 (societal) subjects informed the models. Due to a lack of standard short-term methods for this application, we utilized the visual analog scale (VAS). Waiting trade-off (WTO) tolls provided an additional option for linear transformation of the TMI. We randomized participants to one of three surveys: the first derived weights for generic testing morbidity attributes and levels of severity with the VAS; a second developed VAS values and WTO tolls for linear transformation of the TMI to a death-healthy scale; the third addressed initial validation in a specific test (breast biopsy). 188 patients and 425 community subjects participated in initial validation, comparing direct VAS and WTO values to the TMI. Alternative TMI scoring as a non-preference summated scale was included, given evidence of construct and content validity.

Results—The patient model can use an additive function, while the societal model is multiplicative. Direct VAS and the VAS-scaled TMI were correlated across modeling groups ($r=0.45$ to 0.62) and agreement was comparable to the value function validation of the Health Utilities Index 2. Mean Absolute Difference (MAD) calculations showed a range of 0.07 – 0.10 in patients and 0.11 – 0.17 in subjects. MAD for direct WTO tolls compared to the WTO-scaled TMI varied closely around one quality-adjusted life day.

Conclusions—The TMI shows initial promise in measuring short-term testing-related health states.

Preference-based health-related quality of life (HrQoL) assessments, whether they use direct measures^{1,2} or indexes^{3–7}, focus on measuring longer-term health states. However, shorter duration states such as diagnostic testing are commonly encountered, which may influence long-term outcomes⁸. Direct preferences have limitations for valuing, for example, the morbidity of colorectal screening, a known barrier to adherence^{9,10}. Limitations include choices that juxtapose short-term health states and death. A likely result is that the short-term state is trivialized and little time is traded using time trade-off (TTO) or risk taken with Standard Gamble (SG)¹⁰. A solution for such assessments has been a two-stage chained method, which distances the temporary state from death. In the first stage, an anchor state is valued between dead and perfect health, and in the second, the temporary health state of interest is valued between the anchor and perfect health. This approach is not useful for testing, because with temporary TTO it is still difficult to frame the choices. One must value tests using varying lengths of time in the testing health state, which is problematic, since we do not experience a test over a period of months or years.^{11, 12} Attempts to adapt current methods to testing and screening disutility include Willingness to Give up Time¹³, visual analog scale (VAS) adaptations^{14, 15} and the Waiting Trade-off (WTO) from TTO^{15–17}.

The waiting trade-off (WTO) addresses temporary TTO issues by presenting a choice between having a potentially noxious test of interest, as opposed to a hypothetical “ideal” test with no pain or risk. The noxious test choice allows for immediate results and treatment, but the ideal test requires waiting for the results. The assumption of the WTO is that individuals will wait more to avoid a noxious test than one that is benign.

Appendix A addresses the calculation of a WTO toll and relationship of the WTO to temporary TTO as a variant of a chained method. Among generic indexes, only the EQ-5D has been used for testing^{18, 19}. A known limitation in using the EQ-5D in short-term health states is its pronounced ceiling effect²⁰.

An index capable of measuring the effects of any diagnostic test would improve HrQoL assessment. A parsimonious, face-valid classification of diagnostic testing domains has been described, now called the Testing Morbidities Index (TMI)²¹. The TMI consists of seven attributes for physical and mental aspects of HrQoL prior, during, and after testing (Table 1). Prior psychometric analysis suggests content and construct validity of the 7 TMI items²¹. The TMI might therefore be applicable in preference-based or summated scale versions (Appendix B), and scaling alternatives could allow varied applications. A summated scale may be more responsive than a preference-based instrument^{22, 23}.

We describe derivation of patient and societal preference functions for the TMI by adapting multiattribute value methods^{3, 24, 25}. We applied the TMI initially in breast biopsy, comparing it to direct VAS and WTO. Secondary goals included exploration of model specification and alternative scoring as a summated scale.

METHODS

Ethics

The Partners Healthcare/Massachusetts General Hospital Institutional Review Board (#2005P002377) approved the study.

Overview of Participants and Rationale

The societal perspective is often preferred for HrQoL data in cost-effectiveness analysis, but its advantages over a patient perspective remain controversial^{1, 26, 27}. An instrument using the patient perspective is the Patient-Oriented Prostate Utility Scale (PORPUS)²⁵. The patient and societal perspectives are both relevant for the TMI, since either view can affect testing adherence. Any reference to “subjects” or “community” refers to our societal sample. “Participants” refers to patients and subjects. Both societal and patient groups had survey items assessing numeracy and literacy and provided socioeconomic data (Table 2)^{28, 29}.

Patient Perspective Group—To obtain an estimate of preferences in patients who recently experienced a test, we recruited English and Spanish-speaking women who had breast biopsy. Demographics of patients seen in our breast center are representative of our U.S. Census region³⁰.

On the date of biopsy, a study coordinator provided patients with survey materials. A VAS feeling thermometer was included with a letter that introduced the study and described assessments. Telephone interviews occurred within 3 days after biopsy (average) and before results were known, allowing patients to review the survey and focus on the testing experience without the influence of knowing the biopsy results. Mixed-mode survey approaches including telephone have been used successfully^{15–17, 31}.

Subject (Societal) Perspective Group—We utilized our patient registry that includes over 1.5 million people³². Eligible subjects were ages 21–80, of either gender, and spoke English or Spanish. We excluded those with previous breast biopsy and other recent testing. We drew random samples within strata of age group, gender, race/ethnicity on a quarterly basis to assure subject assessment in the approximate timeframe of patient accrual. We mailed cover letters and survey instruments to subjects and followed up with telephone interviews.

Modeling Approach: Multiattribute Value Theory

Utility index development utilizes either multiattribute utility, multiattribute value theory^{3, 24} (MAUT/MAVT), or multiple regression^{4, 5}. Recent work from the National Health Measurement Study shows reliability coefficients for the MAUT/MAVT-based Health Utilities index Mark 2 and 3 (HUI2/3)³³ that are similar to or better than regression-based indexes. Larger sample sizes are generally needed for regression models compared to MAUT^{25, 34}, and we planned to develop two separate value functions to capture the patient and societal perspective, requiring two modeling groups. Finally, ceiling or floor effects have been substantial in some regression models, which might exacerbate the already likely potential for such problems in the TMI, regardless of method^{35, 36}. Given these issues, and because we were not assessing risky outcomes, we used MAVT.

MAVT extends utility theory to explicit multiple HrQoL domains or attributes^{37, 38}. The HUI2/3 is the most widely known MAUT/MAVT-based index. The HUI instruments have the multiplicative form, shown in EQ.1 in Table 3. The details of MAVT calculations are discussed in Appendix C.

Modeling Groups Overview (Figure 1)

Our approach had three main goals: 1) Derive weights for the TMI classification and attributes, 2) Derive preferences to enable rescaling of the TMI multiattribute valuation model on a dead-healthy scale, and, 3) Collect data for a first estimate of validation. To accomplish these tasks, three survey groups were set up identically for patients and subjects with randomized group assignment. For the first goal, an “attribute valuation group” valued

the TMI attributes and their levels considering the effects of testing situations in general. For the second goal, an “index scaling group” valued the best and worst possible TMI health states for rescaling the index. Based on pilot studies, assessments for the first and second goals utilized a mix of disvalue and value perspectives. These differences required reconciliation with EQ. 2 as shown in Fig. 1 and Table 3. Calculations used software modified from a prior application³⁹. We addressed the third goal with a “validation group” comparing the TMI to directly elicited WTO and VAS, using breast biopsy as a first model test. Patient sample sizes were based on prior published WTO data¹⁵, given its greater variability than VAS. Societal group sizes were based on group preferences likely having a Standard Deviation (SD) 1.5 times greater than patients⁴⁰.

The attribute valuation and index scaling groups were given graphics and text descriptions of tests of varying invasiveness (colonoscopy, phlebotomy, breast biopsy and chest radiography). The validation group received the breast biopsy description.

Attribute Valuation Group: Attribute Levels, Weights and Scaling Constant

TMI Attribute Levels—Participants assigned a value to these levels, bearing in mind the likely impact of varying the level of each attribute when considering any (or many possible) diagnostic tests, assuming normal status of other attributes. Participants were presented with 5 levels within each attribute using a disvalue framework, where “none” was given a default value equaling 0 (no disvalue) and “extreme” a default value equaling 100 (“Attribute Levels” in Fig. 1), later converted to a 0 to 1.0 scale. Participants gave their own VAS values for the “mild”, “moderate” and “severe” levels^{24, 25} (Table 1).

TMI Attribute Weights (k_j) and Scaling Constant (K) from EQ. 1—Participants rank-ordered attributes by importance in a first “warm-up” step. Assignment of weights (see “Attribute Weights” in Fig. 1) was adapted from a combination of MAVT techniques³⁸. Participants assigned 10 points to the least important attribute, and points greater than 10 to all others using ratios, by relative importance (importance weights) in this second step. We obtained the VAS value for the “corner state” of the most important attribute in our third step. A corner state in a disvalue framework is a health state where an attribute is at its worst level, with all other attributes at their best (all other attributes assumed to be normal or disutility of zero). The corner state defines an attribute’s weight (k_j). We used the one corner state with the importance weights to solve for other corner states (k_j) in the final step to minimize burden, instead of directly obtaining all corner states. VAS valued the corner states with “best” and “worst imaginable health state” scale anchors⁴ at 100 and 0, respectively, thus a value framework, requiring reconciliation to disvalue model terms like the levels (Fig. 1 and EQ. 2 in Table 3). Using s_j to denote the importance weight of the j^{th} attribute and the most important attribute corner state preference as $v^{(1)}$, s_j is a “relative” $v^{(j)}$ ³⁸. The ratio of s_j/s_1 is approximately equal to the ratio of $v^{(j)}/v^{(1)}$. Since $v^{(1)}$ is obtained directly and:

$$k_1 = v^{(1)} \quad (EQ. 3)$$

the other attribute weights were derived, thus:

$$k_j = k_1 s_j / s_1 \quad (EQ. 4)$$

The scaling constant K is iteratively calculated from the weights and EQ. 1 as in Appendix C. Therefore, the TMI model developed variables in disvalue terms, then conversion to a value function occurred (EQ. 2) in anticipation of transformation to a dead to perfect health scale (Fig. 1).

Index Scaling Group: Derivation of TMI Rescaling Preferences

Because attribute valuation groups provided patient and societal TMI models on a “worst” to “best” 0 – 1.0 scale (called “final TMI value models” in Fig. 1), conversion to a “dead” to “perfect health” scale required linear transformation using preferences for the best (“*TMIbest*”) and worst (“*TMIworst*”) possible TMI states obtained from the index scaling group. We elicited VAS values for *TMIbest* and *TMIworst*, and due to a lack of applicable standard choice-based measures, we also utilized WTO tolls.

VAS Valuation—The *TMIworst* multiattribute health state summarized a hypothetical test with all worst levels in all attributes using the TMI classification language. Each participant chose *TMIworst* as valued greater than, equal to, or worse than dead. If equal to dead was chosen, then the individual was classified as operating on a 0 equals dead and 100 equals perfect health scale. If *TMIworst* was valued greater than dead on the VAS, then on the VAS scale, again 0 equals dead, and *TMIworst* valuation was done accordingly. If *TMIworst* was considered worse than dead, then dead was valued above *TMIworst* on a 0 – 100 (Perfect Health) VAS. Subsequent linear transformation of *TMIworst* in this latter group followed the work of Patrick, where the worst state, $V(x)$, is rescaled as:

$$[(V(x) - V(\text{dead})) / (1 - V(\text{dead}))] \quad (\text{EQ. 5})$$

allowing negative preferences, with -1.0 the lowest allowed value⁴¹. Group trimmed means for *TMIworst* were weighted by prevalence of those valuing it greater than or equal to dead vs. those valuing it worse than dead, similar to the HUI3^{3, 42}. The *TMIbest* state was valued similarly to *TMIworst*, using a multiattribute TMI scenario with all best levels. For *TMIbest* VAS valuation, the top anchor was at perfect health and the bottom at the worst imaginable state or dead, depending on the prior assessment of *TMIworst*.

WTO Valuation—The WTO derives a quality-adjusted life year (QALY) toll estimate by defining an indifference point between waiting for results of a hypothetical “ideal” test (brief, with no risk or pain) and treatment, and having an actual, potentially noxious test where results and treatment are available immediately^{15–17}. Waiting time was quality-adjusted by baseline and test-related anxiety preferences, requiring multiplication by 0.2 (Appendix A). To minimize anchoring bias, initial waiting time offers for *TMIworst* were randomized between 2, 4, 6 and 8 weeks. We offered 1 day if the participant declined the initial offer, and the assessment stopped if the individual declined 1 day (assumed indifference point of 0.5 day). Otherwise, bisection of the interval between the initial offer and 1 day occurred until indifference, unless the initial offer was accepted. In that case, 2 week intervals were added until a “no” response, followed by bisection to indifference¹⁵. *TMIbest* was similarly valued.

Validation Groups A and B: Comparing the TMI with Direct Preferences

Subjects and patients provided preferences for breast biopsy with direct VAS, direct WTO and TMI items. For subjects, assessments used slightly altered hypothetical language (e.g., for the TMI pain item during a test, “...how much pain or discomfort *do you think you would* experience...” instead of “...how much ...*did* you experience...”)²¹. Group “A” had VAS anchors of “dead” and “perfect health” with sample size assumptions similar to the attribute valuation and index scaling groups, while Group “B” had direct VAS anchors of “worst” and “best imaginable health state”. These groups were developed due to bias concerns when “dead” was the lower scale anchor in patients⁴³. Group “B” began early in the study, with sample sizes based on interim data.

Statistical and Other Modeling Considerations

Descriptive Statistics for Modeling and Agreement—We utilized three methods from HUI development^{3, 24,40}. First, a “person mean” was used, where instead of computing individual participants’ indexes, overall means for each variable (weights for k_j variables, level values and *TMIbest/TMIworst* preferences) were obtained. The constant K and TMI model were subsequently calculated (Appendix C). Second, VAS data for model development were 10% trimmed means (Table 3); 10% trimmed means have also been advocated in the statistical literature⁴⁴. Third, mean absolute difference (MAD) and Overall Standard Deviation (OSD) calculations compared agreement of direct VAS or WTO with the VAS or WTO-scaled TMI (Table 4, Appendix C). The OSD is similar to a root mean square error calculation used in the PORPUS²⁵, except that, like the HUI3 formula, our OSD utilized trimmed means of direct measures instead of arithmetic means. For additional measures of agreement, mountain plots and Pearson correlations were calculated, supplemented with Spearman correlations.

Given no standards for analyzing WTO descriptive statistics, we calculated arithmetic, 5% and 10% trimmed means for the *TMIworst* and *TMIbest* values (Appendix C), and both of these trimmed means served to rescale the TMI calculations to WTO. We calculated arithmetic and 10% trimmed means for the direct WTO tolls (Table 4).

To estimate ceiling or floor effects, we calculated the percentage of each validation subgroup at the top or bottom of the TMI scale, respectively.

Mann-Whitney U tests compared patient and societal preferences. Spearman correlation coefficients compared rank-ordered attributes with importance weights from the attribute valuation group, as a check on importance weighting. Analyses utilized MedCalc v. 11.4.3 (Mariakerke, Be.), Microsoft Excel 2007–10 (Redmond, WA.) and SPSS v. 16.0. (Chicago, Ill.).

The TMI as a Summated Scale and Further Exploration of the TMI Model—

Appendix B explores psychometric evidence and scoring of the TMI as a summated scale. In Appendix C, we include details on MAVT, additive vs. multiplicative models, patient and subject model differences, worse than dead states, scaling parameters and agreement formulae. Appendix D summarizes TMI results by quartile of VAS scores. A downloadable version of the TMI models supporting calculations is available from the authors’ website at http://www.mgh-ita.org/TMIv1_1_0.xls.

Results

Patients and Subjects: Demographics and Response Rates

Twelve hundred eighty five individuals participated (394 Patients, 891 Subjects, Table 2). Patient response rate was 93% among those contacted during the period of eligibility and 79% among patients at the time of biopsy. Subject response rate was 55%. Ten incomplete surveys were noted, with four yielding usable data (Table 2). Population demographics were consistent with our U.S. Census region³⁰.

On average, 4.6% said they could not perform certain health communications tasks. 73.3–77.2% answered numeracy questions correctly on probability and percent risk, respectively (Table 2); a “highly educated” sample in the literature correctly answered them 78 and 84% of the time²⁸. No numeracy issues were suggested from interviews or overall results.

Attribute Valuation and Index Rescaling Groups

Participants indicated no difficulties in valuing corner states. Comparison of rank ordering of attribute importance and ratio importance weights showed significant Spearman correlations in the patient group (mean ρ : -0.69) and the subjects (mean ρ : -0.73) (range across groups: -0.58 to -0.83 , all $p < 0.001$). The patient model result was weakly multiplicative, with $K_p = 0.03$ (Table 3), and is solvable with the additive equation (Appendix C: EQ. 1C). Attribute final scale weights for patient or societal models were between 0.12 and 0.20, excluding Embarrassment. The societal model was somewhat more multiplicative in structure ($K_s = 0.27$, Table 3, Appendix C).

Similar to the PORPUS attribute level valuations²⁵, our least important attribute, embarrassment, showed less mean disutility in more severe levels. There was acceptable to good internal consistency of the TMI items in the validation group (Appendix B).

Index scaling group societal/patient values (Appendix C) with VAS showed 10% trimmed mean *TMI_{worst}* values at 0.31/0.39 and *TMI_{best}* at 0.90/0.93, respectively. Thirteen subjects and 3 patients contributed worse than dead values to the *TMI_{worst}* weighted results. Allowing health states worse than dead with VAS did not affect the models (Appendix C). WTO waiting time values for *TMI_{worst}* were more variable, unlike *TMI_{best}* WTO results.

Validation Groups

The 4 Survey groups were kept separate for comparisons (Table 4), due to scale anchors for VAS and *TMI^{VAS}*, and to keep those with identical surveys and perspectives (patient or subject) together. Patient groups trended similarly as did subject groups. Direct VAS for biopsy (Table 4) indicated greater disutility in subjects (0.64–0.70 trimmed mean) compared to patients (0.79), which was significant. Table 4 shows Validation “A” subjects ($n=235$) with two extreme outliers excluded from direct WTO tolls, who waited over 6 and 12 thousand unadjusted days. The mean/Standard Deviation (SD) WTO toll in quality-adjusted life days (QALD) with these outliers excluded, as in the table, was 2.80/3.84, as opposed to 19.24/185.18 with them. Direct WTO showed a similar trend to VAS in its 10% trimmed mean values, with a greater toll in subject groups than patients had, but this difference was not significant.

The TMI models also showed significantly increased disutility in subjects vs. patients (increased WTO tolls and decreased VAS). Patient and subject groups trended similarly.

Agreement appeared better in patients than subjects overall, with MAD and OSD measures smaller in patients with the VAS and WTO scaled models. Similarly, mountain plots (Fig. 2A and B) showed less bias in patients. However, both groups showed long tails in the distributions indicating suboptimal agreement, usually due to lower VAS values. There was a marked tendency for participants to anchor on 0.05–0.1 increments on the 0–1.0 VAS scale (Fig. 3A–B).

TMI^{WTO} results in Table 4 for validation group “A” exclude the two subject outliers. Small differences in MAD or OSD occurred when including them (range: 0.02 – 0.07 QALD).

There were moderate to strong^{45, 46} Pearson correlation coefficients between direct VAS and *TMI^{VAS}* across validation group patients and subjects ($r=0.45$ to 0.62 , all $p < 0.0001$). For direct WTO vs. *TMI^{WTO}*, one group (“B” patients) was significant at the 0.05 level ($r=0.31$, $p=0.0039$). Group A and B subjects showed weak Spearman correlations between WTO and *TMI^{WTO}* ($\rho = 0.16$ – 0.17 , $p = 0.0095$ – 0.02).

There was no TMI floor effect. The ceiling effect in Group “A” and “B” patients and subjects was 6%, 1.7%, 3.4% and 0.5%, respectively.

Discussion

We applied preference weights to a health classification for measuring short-term HrQoL related to diagnostic testing. The TMI has 5 different scaling alternatives: 1). a patient perspective TMI^{VAS} on a dead-healthy scale; 2). a patient TMI^{WTO} where the worst-best VAS model is rescaled with quality-adjusted WTO data for the best and worst model states; 3). and 4)., the two prior options from the societal perspective; and 5). a summated scale version. Therefore, the TMI could find applications in patient or subject groups, or where preferences are not needed.

We followed the general techniques used in the development of other MAUT-based indexes, where a health classification is valued on a variant of a “best”-“worst” scale, followed by linear transformation to a “dead”-“perfect health” scale^{3, 25}. However, in developing the TMI, the choices made had strengths and limitations. We will discuss our choices and results in 6 categories below.

Multiatribute Value Theory

As cited in the methods section, MAVT provided multiple advantages. Temporal separation of TMI attributes (before, during, after) likely aided in mutual preferential independence (called “structural independence” by the HUI Group⁴²), and we had successful corner state valuations. However, in order that a MAVT-based instrument is generic, attributes must be valued in a generic sense. Thus, participants valued attributes considering the impact of any test experience. This method is similar to a typical generic index where a health state is valued with varying levels of the attributes, not otherwise specified as to disease or other context. Such an exercise involves a significant change in focus for some people, though preference assessments in general are not usual activities for most of us. To what degree such assessments are truly understandable in a generic context has not been explored, but the fact that existing indexes tend to correlate well with one another suggests meaningfulness of their constructs⁴⁷.

Structural independence in MAVT has long been a concern and has partially motivated the development of statistical modeling for indexes⁵. Such independence assumes that attributes can be valued independent of one another and that corner states can be valued and serve as attribute weights. Nonetheless, development of independent attributes is often challenging. Therefore, the plausibility of corner states is a potential limitation to applying MAVT in any model.

Direct Measures

Though using the VAS is controversial^{43, 48}, its validity has been advocated recently⁴⁹. It is included in two widely used instruments; the Quality of Well Being Scale⁶ and HUI2/3, in the latter transformed to a Standard Gamble (SG)^{3, 24}. In the HUI3, a 10% trimmed group mean for VAS as well as SG utilities reduced the influence of outliers on the mean estimate. In our use of the 10% trimmed mean VAS, we found expected trends, where temporary health states were valued as being low, but usually above death. Even so, limited feasible direct measures created challenges. The VAS is most feasible, but with no standard choice-based measure to which VAS can be transformed, our choice was to use the WTO as an option for rescaling the index. Adding to these issues is the tendency for individuals to anchor on coarse VAS increments (0.05–0.10 on a 0–1.0 scale, Fig. 3A–B). Similar findings were noted by Raisch et al. in recent work in which a similar “feeling thermometer” VAS

was used alongside the SF-6D and HUI2/3⁵⁰. Otherwise, we found no apparent trend by varying the VAS scale anchors in validation groups “A” vs. “B”. This suggests that participants focused more on the scale rather than the anchors, which may be most relatable to a testing state.

The WTO avoids the infeasibility of usual choice-based measures in testing scenarios. Prior work has suggested evidence of WTO validity^{15–17}. Though the WTO is conceptually equivalent to the chained temporary TTO¹⁵ (Appendix A), its open-ended scale could allow indefinite waiting. Further, reasons for waiting may be complex. Given the WTO’s relative novelty, there is no guidance as to how to handle outliers. We presented the TMI^{WTO} where the initial model is rescaled with 5 or 10%-trimmed means of TMI_{best} and TMI_{worst} direct WTO values. Mean untrimmed direct WTO QALD tolls in our study were consistent with published values for breast biopsy¹⁵, and face-valid WTO wait times were found for TMI_{best} and TMI_{worst} states (Appendix C), but trimmed mean estimates may be more meaningful. Though variability of the direct WTO was noted (Table 4), the MAD and OSD for TMI^{WTO} vs. WTO at close to 1 QALD across groups was promising.

Agreement and other aspects of performance

There was evidence of similar agreement of TMI^{VAS} with direct measures when comparing to the original HUI2 validation study⁵¹. Specifically, the HUI2 MAD was 0.08 for its person-mean disvalue function²⁴; our VAS MAD was 0.07–0.10 in three out of four groups with one at 0.17. However, when the HUI2 multiattribute VAS disvalue function result was used to derive the multiattribute disutility function from the power curve calculation, the MAD was 0.15, thus, close to our worst group at 0.17. Further, the HUI2 OSD was 0.10 and 0.23 for the person mean disvalue function and the disutility function derived from the multiattribute disvalue function, and ours ranges over 0.08 to 0.19 for TMI^{VAS} vs. direct VAS^{24, 51}. The HUI2 has good reliability coefficients in recent literature³³. These results are also encouraging since variability is likely greater in our scenario, where each participant’s ex ante or ex post preference is being calculated. In contrast, identical marker states were valued ex ante by all HUI2 subjects.

While our agreement results are no worse than HUI2 vs. its VAS, disagreement is substantial in some individuals (Fig. 2A–B). Arnold et al.,⁵² has shown that agreement between direct measures and indexes is suboptimal. In many cases, the lack of agreement between indexes and direct measures might be explainable by the subject perspective of indexes as opposed to the patient perspective in direct assessment. However, Arnold also noted this bias when the perspective issue is not relevant - when “hypothetical patients” are giving their preferences. Findings of lack of agreement are also noted across indexes in the work of the National Health Measurement Study, which has shown that though utility indexes share constructs such as “physical”, “psychosocial” and “pain”⁵³, and correlate with one another⁴⁷, they are not interchangeable in terms of agreement⁵⁴.

Encouraging results came from our observations on worse than death states and ceiling effects. Worse than death states were not a substantive influence as expected. Ceiling effects were smaller than or comparable to the best performing generic indexes in this respect. EQ-5D ceiling effects range from 36 to 49%, the HUI2/3 from 4.5 to 15.4% and SF-6D from 1.7 to 5.8%^{20, 33, 36}.

Patient and Subject Perspectives

Patient group results were generally similar to one another as were societal groups across measures (Table 4). Patient data in men are limited by the uncommon need for breast biopsy in men. Because men tend to value their HrQoL higher than women, their inclusion could

have caused an even greater difference between the patient and societal models⁵⁵. The patient and societal models do appear substantively different (Appendix C), and patients valued quality-adjusted states higher than subjects, consistent with prior literature^{1, 2, 56}. Therefore, even though the modeling groups were valuing attributes of testing in a general context, and subjects had information on testing, the patient relative to subject difference persisted. Among reasons for this difference might be that patients valuing health states have a broader perspective of testing than subjects, who may tend to focus on morbidity²⁷. Mountain plot results (Fig. 2A–B) may suggest that such assessments are more meaningful for patients, given less bias. The appropriate perspective in short-term assessments is unclear at this point, and given that everyone's perspective is relevant for testing adherence, continued evaluation of both points of view seems reasonable.

Respondent Burden and Attribute Valuation

The classic literature in MAVT provides help in decreasing burden. As VonWinterfeldt and Edwards noted³⁸, “in principle it is possible to combine virtually any technique for single-attribute value or utility measurement with any weighting procedure and aggregation model”. In our case, valuation of a single corner state with iterative calculations of other corner states by importance weights likely facilitated a consistent weighting structure in each person, while decreasing burden. Separate randomly chosen groups valued rescaling states (index scaling group) vs. the attribute weights and levels (attribute valuation group), further decreasing burden.

However, the conceptual basis of attribute weighting presented challenges, given our burden concerns. Approaches that utilize importance of an attribute are listed in a complementary fashion with those that do not rely on importance (swing weights or corner states) in the MAVT literature³⁸. Further, adding to the confusion, a discussion of swing weighting may use the word importance in describing its methodology⁵⁷. The main concern is potential conflating of importance and value by respondents. Thus, an attribute that has low weight could receive an inappropriately wide range of level values, similar to an attribute that is highly weighted. While we chose to use importance weights to minimize the number of direct corner assessments, we found reasonable ranges of valued levels (mild to severe) given the resulting attribute weights. Referring to Table 3, in subjects, the attribute with the highest weight had the greatest severe level value (pain during a test at 0.76). Most others fell in the middle with severe levels in the low 0.70's to high 0.60's and similar weights. Mental aspects after testing came next with a severe level of 0.63 and slightly less weight, and the lowest weight for embarrassment is consistent with a severe level of 0.57. In patients, the attributes with highest weights have higher severe levels of 0.75–0.77. The middle group has severe levels in the high 0.60's to low 0.70s (fear during and mental after test) and smaller weights. Embarrassment has a severe level of 0.50 and has a low weight. The one exception we see in patients is that for Pain before the Test, the severe level seems somewhat overvalued at a trimmed mean of 0.79, which would likely have minimal model effects.

Conceptualization of Testing Health States as Process Utility and Limitations of Classification

The conceptualization of the TMI is consistent with the WTO concept, since the disutility/disvalue of testing represents a toll on the process of medical care^{15, 17, 58}. This approach is somewhat different from a typical HrQoL method, where specific times are traded-off or evaluated under uncertainty. WTO scaling for TMI allows a toll on the testing experience deductible from QALE. Further, modelers can attach a desired time frame to the TMI^{VAS}. If such short-term outcomes are not sensitive enough for typical cost-effectiveness analysis

modeling scenarios, they may be useful for patient-centered or comparative effectiveness applications.

The classification system of the TMI may be limited by the response set used, since the degree of morbidity is generally “none” to “extreme”. However, the TMI principally addresses the domains of pain and fear/anxiety, and these constructs are handled similarly in the standard indexes. We have tried to adjust for this situation by giving more context-specific description in the survey item stems. Further, the classification is enhanced by these domains being sampled at three intervals in the testing sequence.

Conclusion

The TMI is a brief 7-item instrument for measuring short-term quality of life in testing from both societal and patient perspectives, with multiple potential scaling options. This study adds to the construct validity information in prior work on the TMI²¹. Future evaluation includes assessment of test-retest reliability, responsiveness in testing scenarios compared to existing indexes, and evaluations over a range of test invasiveness.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Financial support for this study was provided in part by grants from The American Cancer Society 114130-RSGHP-07-266-01-CPHPS, and the MGH ECOR Bridge Fund 1200-218421. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

We acknowledge the assistance of Beverly Gerade, N.P., and Katy Marttila, B.A. This study was funded by American Cancer Society 114130-RSGHP-07-266-01-CPHPS and the Massachusetts General Hospital Bridge Fund ECOR1200-218421.

References

1. Gold, MR.; Patrick, DL. Identifying and valuing outcomes. In: Torrance, GW.; Fryback, DG.; Hadorn, DC.; Kamlet, MS.; Gold, MR.; Siegel, JE.; Russell, LB.; Weinstein, MC., et al., editors. Cost-effectiveness in health and medicine. New York: Oxford; 1996. p. 82-134.
2. Froberg DG, Kane RL. Methodology for measuring health-state preferences--II: Scaling methods. *J Clin Epidemiol.* 1989; 42:459–71. [PubMed: 2732774]
3. Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. *Med Care.* 2002; 40:113–28. [PubMed: 11802084]
4. Kind, P. Chapter 22: The EuroQoL Instrument: an index of health-related quality of life. In: Spilker, B., editor. *Quality of Life and Pharmacoeconomics in Clinical Trials*. 2. Philadelphia: Lippincott-Raven; 1996.
5. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. *J Health Econ.* 2002; 21:271–92. [PubMed: 11939242]
6. Kaplan RM, Anderson JP. A general health policy model: update and applications. *Health Serv Res.* 1988; 23:203–35. [PubMed: 3384669]
7. Hawthorne G, Richardson J, Osborne R. The Assessment of Quality of Life (AQoL) instrument: a psychometric measure of health-related quality of life. *Qual Life Res.* 1999; 8:209–24. [PubMed: 10472152]
8. Cullen J, Schwartz MD, Lawrence WF, Selby JV, Mandelblatt JS. Short-term impact of cancer prevention and screening activities on quality of life. *J Clin Oncol.* 2004; 22:943–52. [PubMed: 14990651]

9. Rawl S, Champion V, Menon D, Loehrer P Sr, Vance GH, Skinner CS. Validation of scales to measure benefits and barriers to colorectal cancer screening. *J Psychosoc Oncol.* 2001; 19:47–63.
10. Wright DR, Wittenberg E, Swan JS, Miksad RA, Prosser LA. Methods for measuring temporary health States for cost-utility analyses. *Pharmacoeconomics.* 2009; 27:713–23. [PubMed: 19757865]
11. Torrance GW. Measurement of health state utilities for economic appraisal. *J Health Econ.* 1986; 5:1–30. [PubMed: 10311607]
12. Swan JS, Miksad RA. Measuring the quality-of-life effects of diagnostic and screening tests. *J Am Coll Radiol.* 2009; 6:567–75. [PubMed: 19643385]
13. van Helvoort-Postulart D, Dirksen CD, Kessels AG, van Engelshoven JM, Myriam Hunink MG. A comparison between willingness to pay and willingness to give up time. *Eur J Health Econ.* 2009; 10:81–91. [PubMed: 18437436]
14. Obuchowski NA, Lieber ML, Magdenic M, Modic MT, Ruggieri PM, Masaryk TJ. Small but quantifiable patient preference for MRA versus catheter angiography. *Stroke.* 1999; 30:2247–8. [PubMed: 10512994]
15. Swan JS, Lawrence WF, Roy J. Process utility in breast biopsy. *Med Decis Making.* 2006; 26:347–59. [PubMed: 16855124]
16. Swan JS, Fryback DG, Lawrence WF, Sainfort F, Hagenauer ME, Heisey DM. A time-tradeoff method for cost-effectiveness models applied to radiology. *Med Decis Making.* 2000; 20:79–88. [PubMed: 10638540]
17. Swan JS, Sainfort F, Lawrence WF, Kuruchittham V, Kongnakorn T, Heisey DM. Process utility for imaging in cerebrovascular disease. *Acad Radiol.* 2003; 10:266–74. [PubMed: 12643553]
18. Verkooijen HM, Buskens E, Peeters PH, Borel Rinkes IH, de Koning HJ, van Vroonhoven TJ, et al. Diagnosing non-palpable breast disease: short-term impact on quality of life of large-core needle biopsy versus open breast biopsy. *Surg Oncol.* 2002; 10:177–81. [PubMed: 12020672]
19. Domeyer PJ, Sergeantanis TN, Zagouri F, Zografos GC. Health-related quality of life in vacuum-assisted breast biopsy: short-term effects, long-term effects and predictors. *Health Qual Life Outcomes.* 2010; 8:11. [PubMed: 20102642]
20. Luo N, Johnson JA, Shaw JW, Coons SJ. Relative efficiency of the EQ-5D, HUI2, and HUI3 index scores in measuring health burden of chronic medical conditions in a population health survey in the United States. *Med Care.* 2009; 47:53–60. [PubMed: 19106731]
21. Swan JS, Ying J, Stahl J, Kong CY, Moy B, Roy J, et al. Initial development of the Temporary Utilities Index: a multiattribute system for classifying the functional health impact of diagnostic testing. *Qual Life Res.* 2010; 19:401–12. [PubMed: 20084464]
22. Krahn M, Bremner KE, Tomlinson G, Ritvo P, Irvine J, Naglie G. Responsiveness of disease-specific and generic utility instruments in prostate cancer patients. *Qual Life Res.* 2007; 16:509–22. [PubMed: 17091359]
23. Langfitt JT, Vickrey BG, McDermott MP, Messing S, Berg AT, Spencer SS, et al. Validity and responsiveness of generic preference-based HRQOL instruments in chronic epilepsy. *Qual Life Res.* 2006; 15:899–914. [PubMed: 16721649]
24. Torrance GW, Feeny DH, Furlong WJ, Barr RD, Zhang Y, Wang Q. Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2. *Med Care.* 1996; 34:702–22. [PubMed: 8676608]
25. Tomlinson G, Bremner KE, Ritvo P, Naglie G, Krahn MD. Development and Validation of a Utility Weighting Function for the Patient-Oriented Prostate Utility Scale (PORPUS). *Med Decis Making.* 2012; 32:11–30. [PubMed: 21653804]
26. Hunink MG, Glasziou PP. Valuing Outcomes. In: Siegel, JE.; Weeks, JC.; Pliskin, JS.; Elstein, AS.; Weinstein, MC., editors. *Decision Making in Health and Medicine.* Cambridge, UK: Cambridge; 2001. p. 88-127.
27. Ubel PA, Loewenstein G, Jepson C. Whose quality of life? A commentary exploring discrepancies between health state evaluations of patients and the general public. *Qual Life Res.* 2003; 12:599–607. [PubMed: 14516169]
28. Lipkus IM, Samsa G, Rimer BK. General performance on a numeracy scale among highly educated samples. *Med Decis Making.* 2001; 21:37–44. [PubMed: 11206945]

29. Chew LD, Bradley KA, Boyko EJ. Brief questions to identify patients with inadequate health literacy. *Fam Med.* 2004; 36:588–94. [PubMed: 15343421]
30. U. S. Census Bureau. Census. 2000.
31. van Wijck EE, Bosch JL, Hunink MG. Time-tradeoff values and standard-gamble utilities assessed during telephone interviews versus face-to-face interviews. *Med Decis Making.* 1998; 18:400–5. [PubMed: 10372582]
32. Murphy SN, Gainer VS, Chueh H. A Visual Interface Designed for Novice Users to Find Research Patient Cohorts in a Large Biomedical Database. *American Medical Informatics Association, Symposium Supplement.* 2003:489–93.
33. Palta M, Chen HY, Kaplan RM, Feeny D, Cherepanov D, Fryback DG. Standard error of measurement of 5 health utility indexes across the range of health for use in estimating reliability and responsiveness. *Med Decis Making.* 2011; 31:260–9. [PubMed: 20935280]
34. Dolan, P. Modeling the relationship between the description and valuation of health states. In: Murray, CJL.; Salomon, JA.; Mathers, CD.; Lopez, AD., editors. *Summary Measures of population health: Concepts, ethics, measurement and applications.* Geneva: World Health Organization; 2002. p. 501-13.
35. Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. *Health Econ.* 2004; 13:873–84. [PubMed: 15362179]
36. Bharmal M, Thomas J 3rd. Comparing the EQ-5D and the SF-6D descriptive systems to assess their ceiling effects in the US general population. *Value Health.* 2006; 9:262–71. [PubMed: 16903996]
37. Keeney, RL.; Raiffa, H. *Decisions with Multiple Objectives: Preferences and value Tradeoffs.* New York, New York: Cambridge University Press; 1993.
38. von Winterfeldt, D.; Edwards, W. *Decision Analysis and Behavioral Research.* New York: Cambridge University Press; 1986. *Multiattribute Utility Theory: Examples and Techniques;* p. 259-313.
39. Swan JS. A software application for multiattribute utility models. *Acad Radiol.* 2004; 11:190–205. [PubMed: 14974594]
40. Furlong, W.; Feeny, D.; Torrance, GW.; Barr, RD.; Horsman, J. *Guide to Design and Development of Health-State Utility Instrumentation.* McMaster University Centre for Health Economics and Policy Analysis; 1990.
41. Patrick DL, Starks HE, Cain KC, Uhlmann RF, Pearlman RA. Measuring preferences for health states worse than death. *Med Decis Making.* 1994; 14:9–18. [PubMed: 8152361]
42. Furlong W, Feeny D, Torrance GW, Goldsmith CH, DePauw S, Zhu Z, et al. *Multiplicative Multi-attribute Utility Function for HUI3: A Technical Report.* CHEPA Working Paper Series. 1998
43. Torrance GW, Feeny D, Furlong W. Visual analog scales: do they have a role in the measurement of preferences for health states? *Med Decis Making.* 2001; 21:329–34. [PubMed: 11475389]
44. Stigler SM. Do robust estimators work with real data? *Annals of Statistics.* 1977; 5:1055–98.
45. Juniper, EF.; Guyatt, GH.; Jaeschke, R. Chapter 6: How to develop and validate a new health-related quality of life instrument. In: Spilker, B., editor. *Quality of Life and Pharmacoeconomics in Clinical Trials.* Philadelphia: Lippincott-Raven; 1996. p. 49-56.
46. De Vet, HC.; Terwee, CB.; Mokkink, LB.; Knol, DL. *Measurement in Medicine.* New York: Cambridge University Press; 2011. *Validity;* p. 150-201.
47. Fryback DG, Palta M, Cherepanov D, Bolt D, Kim JS. Comparison of 5 health-related quality-of-life indexes using item response theory analysis. *Med Decis Making.* 2010; 30:5–15. [PubMed: 19843961]
48. Brazier J, Green C, McCabe C, Stevens K. Use of visual analog scales in economic evaluation. *Expert Rev Pharmacoecon Outcomes Res.* 2003; 3:293–302. [PubMed: 19807377]
49. Parkin D, Devlin N. Is there a case for using visual analogue scale valuations in cost-utility analysis? *Health Econ.* 2006; 15:653–64. [PubMed: 16498700]
50. Raisch DW, Feeney P, Goff DC Jr, Narayan KV, O'Connor PJ, Zhang P, et al. Baseline comparison of three health utility measures and the feeling thermometer among participants in the action to control cardiovascular risk in diabetes trial. *Cardiovascular Diabetology.* 2012; 11:35. [PubMed: 22515638]

51. Torrance, GW.; Zhang, Y.; Feeny, D.; Furlong, W.; Barr, RD. McMaster University Centre for Health Economics and Policy Analysis Working Paper, #92-18. 1992. Multi-attribute Preference Functions for a Comprehensive Health Status Classification System.
52. Arnold D, Girling A, Stevens A, Lilford R. Comparison of direct and indirect methods of estimating health state utilities for resource allocation: review and empirical analysis. *BMJ*. 2009; 339:b2688. [PubMed: 22128393]
53. Cherepanov D, Palta M, Fryback DG. Underlying dimensions of the five health-related quality-of-life measures used in utility assessment: evidence from the National Health Measurement Study. *Med Care*. 2010; 48:718-25. [PubMed: 20613664]
54. Feeny D, Spritzer K, Hays RD, Liu H, Ganiats TG, Kaplan RM, et al. Agreement about identifying patients who change over time: cautionary results in cataract and heart failure patients. *Med Decis Making*. 2012; 32:273-86. [PubMed: 22009666]
55. Hanmer J, Lawrence WF, Anderson JP, Kaplan RM, Fryback DG. Report of nationally representative values for the noninstitutionalized US adult population for 7 health-related quality-of-life scores. *Med Decis Making*. 2006; 26:391-400. [PubMed: 16855127]
56. Brazier JE, Dixon S, Ratcliffe J. The role of patient preferences in cost-effectiveness analysis: a conflict of values? *Pharmacoeconomics*. 2009; 27:705-12. [PubMed: 19757864]
57. Jia J, Fischer GW, Dyer JS. Attribute Weighting Methods and Decision Quality in the Presence of Response Error: A Simulation Study. *J Behav Dec Making*. 1998; 11:85-105.
58. Donaldson C, Shackley P. Does "process utility" exist? A case study of willingness to pay for laparoscopic cholecystectomy. *Soc Sci Med*. 1997; 44:699-707. [PubMed: 9032837]
59. Stein MB, Roy-Byrne PP, Craske MG, Bystritsky A, Sullivan G, Pyne JM, et al. Functional impact and health utility of anxiety disorders in primary care outpatients. *Med Care*. 2005; 43:1164-70. [PubMed: 16299426]
60. Revicki DA, Brandenburg N, Matza L, Hornbrook MC, Feeny D. Health-related quality of life and utilities in primary-care patients with generalized anxiety disorder. *Qual Life Res*. 2008; 17:1285-94. [PubMed: 18949580]
61. DeVellis, RF. *Scale Development: Theory and Applications*. 3. Thousand Oaks, CA: Sage; 2012.
62. von Winterfeldt, D.; Edwards, W. *Decision Analysis and Behavioral Research*. New York: Cambridge; 1986. Structuring for Decision Analysis; p. 42-3.
63. De Vet, HC.; Terwee, CB.; Mokkink, LB.; Knol, DL. *Measurement in Medicine*. New York: Cambridge University Press; 2011. Concepts, theories and models, and types of measurements; p. 7-26.
64. Fayers PM. Quality-of-life measurement in clinical trials--the impact of causal variables. *J Biopharm Stat*. 2004; 14:155-76. [PubMed: 15027506]
65. Richardson, J. Psychometric Validity and Multiattribute Utility (MAU) Instruments. 2010. [cited Research Paper 2010 (57)]; 1-19]. Available from: www.buseco.monash.edu.au/centres/che/pubs/researchpaper57.pdf
66. Ypma TJ. Historical Development of the Newton-Raphson Method. *SIREV*. 1995; 37:531-51.
67. Shi HY, Chang JK, Wong CY, Wang JW, Tu YK, Chiu HC, et al. Responsiveness and minimal important differences after revision total hip arthroplasty. *BMC Musculoskelet Disord*. 2010; 11:261. [PubMed: 21070675]
68. Coteur G, Feagan B, Keininger DL, Kosinski M. Evaluation of the meaningfulness of health-related quality of life improvements as assessed by the SF-36 and the EQ-5D VAS in patients with active Crohn's disease. *Aliment Pharmacol Ther*. 2009; 29:1032-41. [PubMed: 19222413]
69. Farivar SS, Liu H, Hays RD. Half standard deviation estimate of the minimally important difference in HRQOL scores? *Expert Rev Pharmacoecon Outcomes Res*. 2004; 4:515-23. [PubMed: 19807545]
70. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *J Clin Epidemiol*. 1999; 52:861-73. [PubMed: 10529027]

Appendix A. The Waiting Trade-off

Multiple studies have compared more and less invasive testing methods with the WTO, including MRI angiography vs. catheter angiography in peripheral and cerebrovascular disease^{16,17}, and different modes of breast biopsy¹⁵. Results suggest the general validity of this approach, where greater tolls reflect greater disutility. **Fig. A1** shows t_0 as the time a test begins and t_1 as the time it ends; t_2 is the time the individual is willing to wait to avoid the noxious test, obtained with a typical choice-based search pattern to indifference point. The waiting time ($t_2 - t_1$) is quality-adjusted by the difference between the utility for a relevant baseline health state (U_{base}) and the utility of the relevant disease state (U_{state}). Thus, in **Fig. A1**, the rectangle $abcd$ reflects disutility (disvalue), and:

$$\text{Disutility} = (t_2 - t_1) \times (U_{base} - U_{state})$$

In the case where there are clear clinical symptoms for U_{state} that are due to the disease of interest, the waiting scenario is straightforward. For example, in peripheral vascular disease, a person may wait with symptoms of claudication to avoid having an angiogram by catheterization of a major artery. Waiting with symptoms is a realistic construct. Further, utilities exist in the literature for vascular claudication (U_{state}) as well as treated vascular disease (U_{base})¹⁶. More problematic is the scenario where the patient is asymptomatic from the disease of interest. In such cases (e.g., breast biopsy for cancer or other screening), it has been noted that the individual does have an overt disease state - anxiety, which affects health during the events surrounding testing or screening⁸. TTO values for test-related and baseline anxiety were obtained previously for quality-adjustment of the waiting time in breast biopsy¹⁵, and they were influential in quality adjustment of the WTO for generic use in rescaling the TMI. This prior work showed a mean difference between TTO preferences for test-related and baseline anxiety of 0.17, which was used for quality-adjustment¹⁵. However, larger studies have shown a difference between the mean utility of anxiety disorders⁵⁹ and national baseline norms (mean of male and female norms of 40–49 age group) of 0.195⁵⁵. Therefore, taking these data sources into consideration for a reasonable estimate, waiting times from the initial TMI^{WTO} result and direct WTO assessments in the validation group were quality-adjusted by multiplying them by 0.2 as in the disutility equation above. The final quality-adjusted WTO result is usable as a toll on quality-adjusted life expectancy in economic models. We recognize, however, that this difference between anxiety and baseline may be conservative, given the work of others⁶⁰.

Fig. A2 shows that conceptually, the WTO is a variant of temporary TTO, as has been described¹⁵. “ U_{base} ” (utility of baseline), “ U_{state} ” (utility of the prevailing health state – intermediate in position here) and “ U_{test} ” (utility for the test health state) in the WTO are similar to states we could call “ U_{best} ” (utility for best health state), “ U_{int} ” (utility for intermediate state) and “ U_{worst} ” (utility for the worst health state) in temporary TTO. The main difference in WTO vs. TTO lies in which health state in the chain is varied. In temporary TTO used in a testing context, the actual testing health state length (t_1) is varied (“ U_{worst} ”), while in WTO, the “ U_{state} ” health state (t_2) is varied during assessment (“ U_{int} ” in temporary TTO). Varying the length of t_1 is unrealistic, while varying t_2 is reasonable, representing waiting with symptoms.

Like other adapted TTO techniques, there is a likelihood of proportional trade-off violation in the WTO¹⁰. Other WTO limitations include the potential complexity of waiting as a construct and possible indefinite waiting. To create reasonable bounds on WTO waiting in future applications, the current study data for the TMI_{worst} and TMI_{best} states may be

applicable. Multiple WTO measures of the best and worst testing states are now available, given the arithmetic, 5% and 10% trimmed means in Appendix C.

Appendix B. The Psychometric TMI

A prior study showed evidence of construct and content validity of the TMI health classification²¹. Principal components analysis in that study also showed the TMI items loaded strongly with items from an existing summated scale profile addressing the adherence barriers of various colorectal screening technologies⁹. This expected linkage with barriers was present in patients experienced with colorectal cancer screening as well as societal participants.

In the current dataset, internal consistency of the TMI items was assessed using Cronbach's statistics for the patient and societal validation group participants. A set of survey items should optimally have $\alpha = 0.70$ for use in research group comparisons⁶¹. In the pooled TMI patient groups ($n = 188$), $\alpha = 0.72$, and in pooled societal groups ($n = 425$), $\alpha = 0.80$.

The above lines of evidence taken together suggest the TMI is applicable as a non-preference summated scale. However, it is important to comment on differences in the aims of the conceptual models of utility indexes as opposed to many psychometric scales.

At first consideration, it would seem counterintuitive that a multiattribute value theory-based index could show internal consistency, since independence among attributes may decrease their correlation. From the MAVT perspective, a finding of internal consistency does not necessarily preclude structural independence of a group of MAVT-based attributes, because "environmental correlation" of attributes does not preclude their "judgmental independence" as noted by von Winterfeldt and Edwards⁶². In other words, a group of attributes may show statistical correlation due to circumstances of the application or the groups in which they are applied, while still being viewed as independent for valuation by individuals.

From the psychometric perspective, internal consistency can vary with an instrument's applications, but conceptual frameworks are likely even more important. Utility indexes follow a more "formative" concept, because attributes and survey items together form a construct related to HRQoL. Attributes may vary independently and may be called "causal" variables^{63, 64}. For example, in a generic preference index, if HrQoL decreases, it is not required that all attributes show disutility. Many psychometric instruments, on the other hand, have items that have a "reflective" structure, where the items reflect the underlying HrQoL construct and so are more correlated, often called "indicator" variables or "effect indicators"⁶⁴. For example, indicator variables for psychological distress (e.g., worrying thoughts, restlessness, irritability, etc.) would tend to trend together. Factor analysis and testing of internal consistency assume this reflective structure and may be indeterminate or confusing in formative models. Even so, such conceptual divisions are not always clear; some mixture of formative and reflective aspects may be present. A developer of the AQoL instrument noted these blurred conceptual boundaries, and proposed that while development of the domains of a MAUT-based instrument should look to classical psychometrics, the combination of these same domains into an overall HrQoL concept should use the formative modeling of economics/decision analysis⁶⁵. Since the developmental focus of the TMI was as a utility index, we expect there could be many testing applications where independence aspects of the TMI are much more optimal than its internal consistency.

The above concerns notwithstanding, Fayers has noted that even for a set of formative/causal variables, a simple summated score will frequently be robust⁶⁴. Similarly, Tomlinson et al. and the PORPUS developers also provided summated scoring²⁵. With this in mind, we modified an equation from Tomlinson et al. below to convert the 7-item TMI summated

scale to a 0 to 100 scale, with 100 being the best possible score (best possible test). Each item has equal weight and the response is coded 1–5 in variable L_i . The denominator value of 4 is the number of response categories (5 in all items) – 1.

$$\text{TMI Summated Scale} = 100 - \left[\frac{100}{7} \right] \sum_{i=1}^7 \frac{L_i - 1}{4}$$

Appendix C

1. Multiattribute Value Theory (MAVT)

EQ. 1 from Table 3 in the text is the multiattribute value theory (MAVT) version of a MAUT-based function. As is evident from comparing to EQ. 1A below, MAUT and MAVT preferences use similar formalism^{24, 38}. This appendix presents further details on the use of this and other equations related to it. If EQ.1 is used for a utility function (EQ. 1A), the notations are as follows: $u_j(x_j)$ is the single attribute utility function for an attribute (j), where 0 is the worst morbidity value possible, and 1.0 is the best. The single attribute function reflects utility attached to each of the intermediate levels of function on a 0 to 1.0 scale. K is a global scaling constant that captures the preference relationships among attributes. k_j is the weight applied to each attribute, representing its contribution in the overall utility function. represents multiplication through all attributes $j = 1$ through n ^{3, 37}. EQ. 1B is used for the iterative calculation of K , once the k_j are all known. In our case, after obtaining the group mean k_j value for each attribute, K for patient and societal models was derived with EQ.1B using Newton's Method⁶⁶. If the sum of all k_j is 1.0, K is needed to scale the multiplicative function between 0 and 1.0. In such a case, attributes can only relate to one another as preference complements or substitutes, not both³. If attributes act as complements, the net increase in utility when both improve is greater than if they are substitutes. If all k_j sum to 1.0, a linear additive model is implied ($K = 0$), with no preference relationships among attributes (EQ. 1C).

$$u(x) = (1/K) \left[\prod_{j=1}^n (1 + K k_j u_j(x_j)) - 1 \right] \quad (\text{EQ. 1A})$$

$$\text{where: } (1+K) = \prod_{j=1}^n (1 + K k_j) \quad (\text{EQ. 1B})$$

$$u(x) = \sum_{j=1}^n k_j u_j(x_j) \quad (\text{EQ. 1C})$$

In health scenarios, a disutility (disvalue) model is generally more intuitive, where the absence of morbidity is equal to 0 on the utility scale and 1.0 is equal to the worst health state. Disutility assumptions have been utilized in multiple indexes^{3, 25}. Accordingly, the k_j weight for each attribute is equal to its corner state assessment³, where each attribute is valued at its worst possible level (1.0), with all other attributes assumed to be normal or disutility equal to 0. This approach becomes clear when EQ.1A is set as a disutility model (thus, (x)) for a corner state. In such a case, all other attribute terms drop out (morbidity equal to 0) with disutility equal to k_j for the one attribute at its corner state. Transformation

to a utility (or value model, delineated by $v(x)$) using EQ. 2 is typical: $u(x) = 1 - v(x)$ or, $v(x) = 1 - u(x)$ (EQ. 2 from main text).

MAVT-based indexes assume reasonable structural independence (mutual preferential independence) of the attributes, thus any level of function in one attribute should be conceivable without regard to others³. A set of attributes lacking structural independence cannot generate plausible corner states. Such independence obviates the need for direct valuations of thousands of possible states. In the TMI, temporal separation of attributes (Table 1 in text) facilitates structural independence. Existing indexes show that physical and mental attributes can co-exist in MAUT/MAVT models with reasonable structural independence^{3, 24}.

2. Specification of Additive and Multiplicative Models: Overview

A Ruby script (v. 1.9.2) generated the 78,125 possible seven-element health state vectors in the TMI classification. Each vector provided input for Excel (Microsoft, Inc., Redmond WA) to calculate the model result as well as differences between models. Visual Basic for Applications provided the output of the models in Excel 2010. Because all seven TMI items use the five level response set, the best possible vector is 1111111 and the worst is 5555555. Given that our results showed a variably weak multiplicative model (EQ. 1) in patients and subjects (Table 3), we assessed the differences between our additive and multiplicative models in all possible vectors, as well as the patient vs. societal multiplicative models. Specifically, for each possible health state vector, a MAUT-based additive and multiplicative model calculated preferences using the derived patient attribute weights, level values, scaling constant K for the multiplicative model, and TMI_{best} and TMI_{worst} scaling values (EQ. 1, Table 3 **and heading 5. below**). We restricted modeling comparisons to VAS, since VAS results are more meaningful to readers and the WTO has had limited use. The societal model comparisons were handled similarly. Histograms summarized the results. Similarly, we compared the output of patient and societal multiplicative models.

Additive versus Multiplicative Models – Patients

Given a value of $K_p = 0.03$ (Table 3), it is expected that the additive and multiplicative solutions would be similar in patients. Almost all differences over the 78,125 possible health states are at or less than .01.

Additive versus Multiplicative Models - Societal

Since $K_s = 0.27$ in the societal model (Table 3), more divergence is expected between the additive and multiplicative solutions, though this K value indicates the model is still not strongly multiplicative. Nonetheless, more than 25% of the vector value differences are at 0.02 or greater (**Fig. C1**). The minimal important difference (MID) for the TMI^{VAS} would be helpful to estimate whether these differences between models are important. The TMI MID is not known by anchor-based approaches, but a preliminary estimate is conceivable with distribution-based methods. Others have noted that one-half the standard deviation of the instrument may approximate the MID^{67, 68} in patient-reported outcomes instruments, and in preference-based instruments, 0.3SD has been proposed⁶⁹. Using the preference-based TMI^{VAS} version on a death to healthy scale, our preliminary calculations below suggest the MID may be from .02 – .04. This range is relevant to the societal model comparison above as well as the patient vs. societal model comparison discussed below.

In considering the MID for the TMI, we found that in pooled breast biopsy patients ($n=188$) and pooled societal subjects ($n=425$), the standard deviation of the TMI^{VAS} is 0.062 and 0.073, respectively. This gives a range of 0.018 – 0.022 for 0.3SD estimates across patients

and subjects, and for 0.5SD estimates, 0.031 – 0.037. The Standard Error of Measurement (SEM) ⁷⁰ can also estimate MID, where:

$$\text{SEM} = \text{SD} \sqrt{1 - \text{reliability}}$$

Subject and patient MIDs are 0.031 and 0.033 using these SD estimates and the Cronbach's results from Appendix B for reliability. However, as noted in Appendix B, Cronbach's may be less appropriate in utility models. Test-retest reliability data, once available, may be more optimal for SEM estimates.

3. Patient vs. Societal Modeling Perspective

Differences between the patient and societal multiplicative models appear substantive. The histogram (**Fig. C2**) shows most differences at 0.03 or greater.

4. Health States Worse than Death

We had provided for the possibility of a worse than death health state at 5555555, following usual practices in utility index construction. However, the TMI is unusual in its focus on short-term health states. Our interviews revealed worse than death preferences for health state 5555555 in 3/103 patients and 13/232 subjects in the index scaling group. This result is in contrast to the HUI3⁴², where most subjects in a sample of 256 found the HUI3 worst state or “pits” to be worse than death (233/256). We compared the preferences obtained in the validation group if we used the *TMIworst* state as originally calculated (allowing negative preferences) as opposed to censoring our 3 patient and 13 subject *TMIworst* VAS assessments at 0 (no allowance for negative preferences).

Table C1

TMI with/without Negative Scaling Preferences

Validation “A” Subjects n= 235	Mean TMI ^{VAS} Score
Original model TMI ^{VAS} *	0.78
TMI ^{VAS} with <i>TMIworst</i> censored at zero†	0.79
Validation “B” Subjects n= 190	
Original model TMI ^{VAS}	0.78
TMI ^{VAS} with <i>TMIworst</i> censored at zero	0.78

* The *TMIworst* VAS scaling state at 5555555 (0 – 1.0 death to healthy scale) from the index scaling group was calculated as a 10% trimmed mean. It was weighted by the prevalence of negative (worse than death) preferences in 13 subjects (also a 10% trimmed mean), as noted in the text. *TMIworst* = 0.31 (**heading 5 below**). TMI^{VAS} results in subjects in Table 4 use this *TMIworst* value.

† The *TMIworst* VAS scaling state (0 – 1.0 death to healthy scale) from the index scaling group is a 10% trimmed mean where all preferences originally valued as worse than death are censored at zero. *TMIworst* = 0.33 with this assumption.

Table C1 shows that for the mean TMI score, there is a .01 difference in Validation “A” subjects and no mean difference in Validation “B” subjects, comparing models with and without negative preferences. There was a maximum difference in these models of 0.014 and 0.012 in “A” subjects and “B” subjects, respectively; values at the .01 level or greater only occurred in about 2.5% of each group. The patient groups showed no differences, since the mean value for *TMIworst* in patients remained at 0.39 with or without negative preferences.

5. Model Scaling Parameters and Agreement Formulae

Table C2

VAS and WTO measures for rescaling the TMI

	swa	sba	sw5	sb5	sw10	sb10	pwa	pba	pw5	pb5	pw10	pb10
VAS					0.31	0.90					0.39	0.93
WTO	310.36	1.83	62.06	1.11	30.66	0.81	168.19	1.10	33.84	0.70	31.65	0.63

Notation for Table C2:

p_w, p_b = TMI_{worst} and TMI_{best} values, patients (VAS are weighted means, WTO wait times are unadjusted days)

s_w, s_b = TMI_{worst} and TMI_{best} values, subjects (as above)

$a, 5, 10$ = arithmetic mean, 5% trimmed mean, 10% trimmed mean

v = multiattribute value function as initially calculated on a “best” to “worst” scale

v = multiattribute value function on dead to healthy scale

Conversion to “Dead to Healthy” Scale (example using VAS patient function): $v = v(VAS_{pb} - VAS_{pw}) + VAS_{pw}$

Formulae for Table 4 Comparisons in text (footnote markers refer to Table 4):

Mean Absolute Difference (MAD) and overall standard deviation (OSD) estimates (formulae from Feeny et al. Med Care 2002; 40: 113–128): in Validation “A”, TMI and direct VAS measures are on a death to healthy scale. For Validation “B”, TMI and direct VAS measures are on a worst to best scale.

††MAD VAS: $[(|TMI^{VAS}_{predicted} - 10\% \text{ trimmed mean direct VAS}|)/n]$

‡‡ and §§MAD in QALD toll using 5% or 10% trim mean of scaling states for TMI calculation: $[(|TMI^{WTO}_{predicted} - 5 \text{ or } 10\% \text{ trimmed mean direct WTO}|)/n]$

||OSD: $\sqrt{[\sum (TMI_{predicted} - 10\% \text{ trimmed mean direct VAS})^2 / (n-1)]}$

¶¶ and ***OSD:

$\sqrt{[\sum (TMI_{predicted} - 5\% \text{ or } 10\% \text{ trimmed mean direct WTO toll})^2 / (n-1)]}$

Note: sums are over all members of each group

Appendix D. Mean TMI By VAS Quartile

In addition to analyzing the correlation of direct VAS with the TMI^{VAS} in the validation groups, mean TMI score was calculated by quartile of VAS scores. Mean TMI score increases as VAS increases by quartile.

Quartile VAS	n	Mean TMI “A” Patients	n	Mean TMI “A” Subjects	n	Mean TMI “B” Patients	n	Mean TMI “B” Subjects
1	25	0.787	85	0.743	30	0.758	67	0.711
2	30	0.823	41	0.788	15	0.846	49	0.809
3	31	0.866	57	0.811	24	0.872	42	0.837
4	14	0.890	52	0.821	19	0.874	32	0.860

Fig. 1. Flowchart of TMI Development Plan

The left side of Fig. 1 shows the modeling groups, identically designed for patients and subjects, contributing the attribute levels and weights (attribute valuation group) and scaling preferences (index scaling group) for positive linear transformation of the TMI model. Light grey shaded boxes indicate the perspective of the attribute valuation group participant tasks that were on a worst to best scale, while the index scaling group preferences and the model rescaled by them are on a death to perfect health scale (darker grey boxes). Validation comparisons of direct VAS and WTO values with the respective TMI indexes are shaded as appropriate to their scaling perspective.

Fig. 2.

Fig. 2A. Mountain Plots of Group A Patients and Subjects.

The median difference between TMI and VAS is represented by the curve peak and is very near zero in patients (-0.01). The subjects show a more negative bias, with the median difference at -0.06 . Large negative tails reflecting lack of agreement are seen (left side of plot) due to low VAS values in both groups, greater in subjects. Positive tails, where VAS is greater than TMI, are also noted but are less notable than negative tails, and are more prominent in subjects.

Fig. 2B. Mountain Plots of Group B Patients and Subjects. Bias is greater in patients (median difference -0.045) than Group A and in subjects (-0.14). The mountain distribution is wider in subjects than patients, but both show wide tails, with negative tails more prominent.

Fig. 3A

Fig. 3B

Fig. 3.
Fig. 3A. Cumulative Frequency Distributions: Direct VAS across Validation Groups.

As in Table 4, the subjects gave lower VAS scores for biopsy than patients, but both groups tend to anchor on 0.05 to 0.1 increments on the feeling thermometer scale.

Fig. 3B. Cumulative Frequency Distributions: TMI^{VAS} **across Validation Groups**. Subjects again show lower scores than like-grouped patients. The distributions show a more gradual progression in values than direct VAS as the MAVT-based function increases. The A Group patients and subjects have maximum values of 0.93 and 0.90, respectively, on the 0–1.0 dead-healthy scale, as detailed in Appendix C, while B groups are on a 0–1.0 best-worst scale.

Table 1**Multiattribute Health Classification and Survey Items: Testing Morbidities Index (TMI)**

Attribute	Description of Level	
<i>Pain/Discomfort Preparing for the Test</i> How much pain or discomfort was involved with preparing for the test? This question refers to any diet changes, procedures or medications that may have been required before the actual test.	1	no pain or discomfort (or, not applicable)
	2	mild pain or discomfort
	3	moderate pain or discomfort
	4	severe pain or discomfort
	5	extreme pain or discomfort
<i>Fear/Anxiety Before the Test</i> Based on what you knew about this particular test, how much fear or anxiety did you have about having the test before the test?	1	no fear or anxiety
	2	mild fear or anxiety
	3	moderate fear or anxiety
	4	severe fear or anxiety
	5	extreme fear or anxiety
<i>Pain/Discomfort During the Test</i> Thinking about this test, how much pain or discomfort did you experience while it was happening?	1	no pain or discomfort
	2	mild pain or discomfort
	3	moderate pain or discomfort
	4	severe pain or discomfort
	5	extreme pain or discomfort
<i>Embarrassment During the Test</i> How much embarrassment, if any, did you actually feel during the test experience (for example, revealing parts of your body, movements of the body or other aspects that may have caused you to feel awkward or uncomfortable)?	1	no embarrassment
	2	mild embarrassment
	3	moderate embarrassment
	4	severe embarrassment
	5	extreme embarrassment
<i>Fear/Anxiety During the Test</i> How much fear or anxiety did you feel during this test experience?	1	no fear or anxiety
	2	mild fear or anxiety
	3	moderate fear or anxiety
	4	severe fear or anxiety
	5	extreme fear or anxiety
<i>Mental Function After the Test</i> Thinking about the ability to do your daily activities after this test, how serious were any problems you had with your mental function or state of mind (for example, fear, anxiety, worry, depression, difficulty concentrating, etc.)? This may be difficult for you to separate from the process of being a patient. However as best as you can, try to focus on the particular problems the test itself may have caused.	1	no problems
	2	mild problems
	3	moderate problems
	4	severe problems
	5	extreme problems
<i>Physical Function After the Test</i> Thinking about the ability to do your daily activities after this test, how serious were any temporary problems you had with your physical function, (for example, working, walking, ability to move your back, neck and arms, etc.)?	1	no problems
	2	mild problems
	3	moderate problems
	4	severe problems
	5	extreme problems

Table 2

Demographics of Testing Morbidities Index Development Sample

	Full Sample	Patients	Subjects
<i>Sample Size (N)</i>	1285	394	891
Attribute Valuation Group	337	103	234
Index Scaling Group	335	103	232
Validation Group A	335	100	235
Validation Group B	278	88	190
<i>Age (mean yrs.)</i>	49.3	51.8	48.2
range (yrs.)	21–80	22–80	21–80
<i>Gender (% female)</i>	69.1	100	55.4
<i>Race/Ethnicity (% of group)</i>			
White	84.4	86.0	83.6
Afr. American/Black	5.4	5.9	5.3
Others	10.2	8.1	11.1
Hispanic (%)	7.6	4.8	8.8
Spanish survey (% of Hispanic group)	39.2	6.2	33.0
<i>Education</i>			
Years (mean)	15.5	15.6	15.4
12 years or less (%)	23.8	21.1	25.0
<i>Numeracy queries (% correct)</i>			
Which represents the greatest risk of getting a disease (as a probability: 1 in 100, 1 in 1000, 1 in 10, or “don’t know”)?	73.3	72.6	73.6
Which represents the greatest risk of getting a disease (as a percent: 1%, 10%, 5%, or “don’t know”)?	77.2	75.4	78.0
<i>Medical Literacy (% “always” or “often”)</i>			
Need help reading medical material	6.7	4.5	7.7
Need help filling out forms	4.2	1.8	5.4
Problems learning about one’s condition because of difficulty understanding written information	2.9	0.6	4.0
<i>Response Rate (%:see text details)</i>		93	55
<i>Incompletes</i>			
Overall (usable:unusable)	10(4:6)	5(1:4)	5(3:2)

Table 3

Patient and Subject Multiattribute Functions: Testing Morbidities Index

Attribute	Pain Before		Fear Before		Pain During		Embarrassment		Fear During		Mental After		Physical After	
	V_{s1}	V_{p1}	V_{s2}	V_{p2}	V_{s3}	V_{p3}	V_{s4}	V_{p4}	V_{s5}	V_{p5}	V_{s6}	V_{p6}	V_{s7}	V_{p7}
Levels (Disvalue) 10% (trimmed)														
None 1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mild 2	0.18	0.18	0.19	0.21	0.19	0.19	0.15	0.11	0.20	0.20	0.18	0.18	0.19	0.19
Moderate 3	0.41	0.46	0.41	0.47	0.43	0.45	0.33	0.28	0.42	0.45	0.38	0.42	0.42	0.44
Severe 4	0.73	0.79	0.71	0.76	0.76	0.77	0.57	0.50	0.69	0.73	0.63	0.68	0.72	0.75
Extreme 5	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Importance weight 10% (trimmed)	40	36	41	45	56	55	20	12	40	40	35	39	44	46
Scale weight (k_p or k_s)	0.13	0.13	0.13	0.16	0.18	0.20	0.07	0.05	0.13	0.14	0.12	0.14	0.14	0.16

Notation and Formulae for Multiplicative Multiattribute Disvalue and Value Functions (EQ. 1 and 2 – see also Appendix C):

- $v(x)$ = multiattribute disvalue function for a health state x . $v(x)$ = multiattribute value function for x .
- $v_j(x)$ = single attribute disvalue function for an attribute j . K = global scaling constant for a value/disvalue function
- k_j = scale weight for an attribute j .

$$\bar{v}(x) = [1/K] \left[\prod_{j=1}^7 (1 + K k_j \bar{v}_j(x_j)) - 1 \right] \quad (\text{EQ. 1})$$

$$\bar{v}(x) = 1 - v(x) \quad (\text{EQ. 2})$$

Notation for Table 3:

- V_{s1} or V_{p1} = disvalue level in subject or patient function, first attribute
- K_s , K_p = global scaling constants for subject or patient models using EQ. 1; $K_s=0.27$ $K_p=0.03$
- k_s , k_p = calculated attribute weight, subject function or patient function in EQ. 1

Table 4

Measures of Biopsy Preference: Direct and TMI Measures

	Validation "A" Patients N= 100	Validation "B" Patients N= 88	Validation "A" Subjects N= 235	Validation "B" Subjects N = 190
<i>Direct VAS</i>				
Mean/10% [§]	0.78/0.79	0.77/0.79	0.69/0.70	0.63/0.64
Median	*0.80	*0.80	0.70	0.70
SD	0.19	0.20	0.21	0.23
<i>Direct WTO toll^{//}</i>				
Mean/10%	3.06/1.78	2.34/1.88	2.80/2.22	2.98/2.30
median	1.6	1.4	1.6	1.2
SD	8.07	3.23	3.84	4.80
<i>TMI^{VAS}[¶]</i>				
Mean	0.84	0.83	0.79	0.79
Median	*0.85	[‡] 0.85	0.80	0.81
SD	0.07	0.11	0.07	0.12
<i>TMI^{WTO}^{**}</i>				
Mean/Median 5%	1.29/1.13 [*]	1.27/1.15 [*]	2.61/2.26	2.79/2.49
Mean/Median10%	1.20/1.05	1.19/1.07 [‡]	1.33/1.16	1.42/1.28
<i>MAD (VAS)^{††}</i>	0.07	0.10	0.10	0.17
<i>MAD (WTO 5%)^{†‡}</i>	0.95	0.95	1.12	1.14
<i>MAD (WTO10%)^{§§}</i>	0.81	0.85	1.05	1.01
<i>OSD (VAS 10%)</i>	0.08	0.12	0.11	0.19
<i>OSD (WTO 5%)^{¶¶}</i>	1.12	1.08	1.51	1.51
<i>OSD (WTO 10%)^{***}</i>	0.95	0.97	1.17	1.14

*
p <0.001,[‡]
p =0.0080,[‡]
p=0.0075, patients vs. similar group subjects, Mann-Whitney U.[§]
10% trimmed mean,^{//}
waiting time from WTO is multiplied by 0.2 for QALD toll (Appendix A).[¶]
TMI VAS-based models: Group A; TMI^{VAS} is rescaled with 10% trimmed VAS means of index scaling group states on a death to healthy scale. Group B; TMI^{VAS}(worst to best scale) model result is presented.^{**}
TMI rescaled with 5 or 10% trimmed means of WTO tolls for index scaling group states and multiplied by 0.2 for quality-adjustment. For Mean Absolute Difference (MAD) and Overall Standard Deviation (OSD) formulae details and footnotes, see Appendix C.