Preference Assessment

Henry Glick

Epi 550

March 30, 2011

QALYs

• QALYs are an economic outcome that combines preferences for both length of survival and its quality into a single measure
  – In many jurisdictions, QALYs represent the recommended outcome of cost-effectiveness analysis [Gold et al., Cost-Effectiveness in Health and Medicine, summary recommendation IV.1, p. 308]

The Question QALYs Answers

• How do we decide how much we should pay for:
  – A therapy that saves fully functional lives/life years
  – A therapy that saves less than fully functional lives/life years (e.g., a drug for heart failure that extends survival, but patients spend the extra time in NYHA class III)
  – A therapy that doesn’t save lives/life years but improves patients’ functioning (e.g., patients with heart failure spend most of their remaining years in NYHA class I instead of NYHA class III)
The Patient-Centered Outcomes Research Institute . . . shall not develop or employ a dollars per quality adjusted life year (or similar measure that discounts the value of a life because of an individual’s disability) as a threshold to establish what type of health care is cost effective or recommended. The Secretary shall not utilize such an adjusted life year (or such a similar measure) as a threshold to determine coverage, reimbursement, or incentive programs under title XVIII

The Patient Protection and Affordable Care Act

QALY Scores

- QALY or preference scores generally range between 0 (death) and 1 (perfect health)
  - For example, a health state with a preference score of 0.8 indicates that a year in that state is worth 0.8 of a year with perfect health
  - There can be states worse than death with preference scores less than 0

Typology of Elicitation Methods

- Assesses or does not assess risk
- Scaling vs choice
- Preference for current health or preference for years of survival
- Direct vs indirect elicitation
- Whose preferences?
Incorporation of Risk Preference

- Methods for the assessment of QALYs differentiated by whether or not they incorporate preference for risk
  - Utility when they do
  - Value when they don't
- Refer to preference assessment, preference scores, or preferences when referring to generic assessment of QALYs
- Reserve the terms utility assessment, utility scores, or utilities and value assessment, values scores, and values for our references to methods that are used to derive utilities or values, respectively

Scaling vs Choice

- Scaling
  - Rating scale, visual analog scale, feeling thermometer
- Choice
  - Standard gamble
  - Time trade-off

Risk and Choice

<table>
<thead>
<tr>
<th>Response Method</th>
<th>Question Framing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Certainty (values)</td>
</tr>
<tr>
<td>Scalining</td>
<td>Rating Scale</td>
</tr>
<tr>
<td></td>
<td>Category Scaling</td>
</tr>
<tr>
<td></td>
<td>Visual Analog</td>
</tr>
<tr>
<td>Choices</td>
<td>Time trade-off</td>
</tr>
<tr>
<td></td>
<td>Paired comparison</td>
</tr>
</tbody>
</table>

From Drummond et al., Methods for Economic Evaluation of Health Care, p. 143
Preference for Current Health or Years of Survival

- Methods of assessment are also differentiated by whether they are measured as a series of valuations of current health or whether they are measured by use of an explicit preference mapping for years of survival and their quality.
- Because most preference assessment is based on a series of valuations of current health, in the following discussion, we address methods for this assessment and do not discuss how we develop a preference mapping.

Direct Elicitation vs Indirect Preference Assessment

- Direct elicitation: Direct rating of preference for health by respondent.
- Indirect preference assessment: Uses instruments which have the respondent directly rate functional status across a variety of domains but derives preference score from a scoring rule.
- “Gold standard”: Although not necessarily feasible, QALYs constructed by use of direct elicitation, which incorporates risk, and accounts for duration of health states.

Whose Preferences?

- The Panel on Cost Effectiveness in Health and Medicine has recommended a reference case analysis that uses community preferences to value health.
- They also recommend a sensitivity analysis that uses preferences of persons with the condition.
Outline
- Prescored health classification instruments
  - EQ-5D
  - HUI2
  - HUI3
- Direct elicitation
  - Standard Gambles
  - Time trade-off
  - Rating scale
- Comparison of Methods

Prescored Health State Classification Instruments
- One of the two dominant approaches for QALY measurement uses prescored health state classification instruments (indirect utility assessment)
- Participants' report their functional status across a variety of domains
- Preference scores are then derived from scoring rules that have usually been developed by use of samples from the general public
- These instruments are considered to satisfy the "community preferences" recommendation of the Panel on Cost-Effectiveness

Direct Elicitation
- The second dominant approach for estimating preference scores directly elicits participants’ preferences for their current health
- Methods include:
  - Standard gamble
  - Time trade-off
  - Rating scales
- When administered to study participants, these methods yield measures of patient preference
Scenarios
• A third approach describes disease scenarios to members of the general public and directly elicits their preferences for these scenarios
• We do not discuss this method below. Rather, in what follows, we describe prescored instruments and direct elicitation methods.

Prescored Instruments
• A number of prescored instruments are currently available for the measurement of preference scores for current health
  – The EuroQol instrument (EQ-5D)
  – The Health Utilities Index Mark 2 (HUI2)
  – The Health Utilities Index Mark 3 (HUI3)
  – The SF-6D
  – The Quality of Well-Being Scale (QWB)
  – The 15D
  – The Disability and Distress Index (DDI)
• Most of these instruments ask participants or their proxies to report on the health status of the patient

The EQ-5D, the HUI2, and the HUI3
• The EQ-5D, the HUI2, and the HUI3 are three of the most commonly used prescored preference assessment instruments
• All three of these instruments share features of ease of use
  – e.g., high completion rates and the ability to be filled out in 5 min or less
• All have been used to assess preferences for a wide variety of diseases
The EuroQol instrument

- The EuroQol instrument is made up of two parts:
  - A health state classification instrument (EQ-5D) and its attendant scoring rule
  - A 100-point visual analog scale
- The latter is a form of direct elicitation

EQ-5D Domains

- The EQ-5D health state classification instrument has 5 domains
  - Mobility
  - Self-care
  - Usual activities
  - Pain/discomfort
  - Anxiety/depression

EQ-5D Levels of Function

- Each domain is defined by 3 levels of function from good to poor
- Generally worded:
  - "I have no problems..."
  - "I have some problems..."
  - "I am unable...."
- The 3 levels for each of the 5 domains can be used to define 243 ($3^5$) health states
EQ-5D Scoring Rule(s)

- The principal scoring rule was developed by Dolan by use of time trade-off responses from a representative sample of 2997 noninstitutionalized adults from England, Scotland, and Wales.
- AHRQ supported the development of a U.S. population-based preference weighting system for EQ-5D through a series of time trade-off (TTO) exercises and statistical imputations (http://www.ahrq.gov/rice/eq5dproj.htm).
- Scoring rules exist for at least 10 additional countries (Szende, Oppe, Devlin eds. EQ-5D Value Sets: Inventory, Comparative Review and User Guide. Springer, 2010).

Dolan Scoring Weights

<table>
<thead>
<tr>
<th>Mobility</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I have no problems walking about</td>
<td>0.000</td>
</tr>
<tr>
<td>2. I have some problems walking about</td>
<td>0.069</td>
</tr>
<tr>
<td>3. I am confined to bed</td>
<td>0.314</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Self-Care</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I have no problems with self-care</td>
<td>0.000</td>
</tr>
<tr>
<td>2. I have some problems washing and dressing myself</td>
<td>0.104</td>
</tr>
<tr>
<td>3. I am unable to wash or dress myself</td>
<td>0.214</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Usual Activities</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I have no problems with performing my usual activities</td>
<td>0.000</td>
</tr>
<tr>
<td>2. I have some problems with performing my usual activities</td>
<td>0.036</td>
</tr>
<tr>
<td>3. I am unable to perform my usual activities</td>
<td>0.094</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pain/Discomfort</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I have no pain or discomfort</td>
<td>0.000</td>
</tr>
<tr>
<td>2. I have some pain or discomfort</td>
<td>0.123</td>
</tr>
<tr>
<td>3. I have extreme pain or discomfort</td>
<td>0.386</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Anxiety/Depression</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I am not anxious or depressed</td>
<td>0.000</td>
</tr>
<tr>
<td>2. I am moderately anxious or depressed</td>
<td>0.071</td>
</tr>
<tr>
<td>3. I am extremely anxious or depressed</td>
<td>0.236</td>
</tr>
</tbody>
</table>

Dolan Scoring

- Scoring formula:
  - If all domains are level 1: 1.000
  - If at least one domain has a score of 2 and no domains have a score of 3 (i.e., worst functioning): 0.929 – sum of scores
  - If one or more domains have a score of 3: 0.65 – sum of scores
Sampling of Other Scoring Rules

<table>
<thead>
<tr>
<th></th>
<th>Dolan</th>
<th>Denmark</th>
<th>Netherlands</th>
<th>Spain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any 2 or 3</td>
<td>-.071</td>
<td>-.114</td>
<td>-.071</td>
<td>-.024</td>
</tr>
<tr>
<td>Any 3</td>
<td>-.279</td>
<td>--</td>
<td>-.234</td>
<td>-.291</td>
</tr>
<tr>
<td>M2</td>
<td>-.069</td>
<td>-.053</td>
<td>-.036</td>
<td>-.106</td>
</tr>
<tr>
<td>M3</td>
<td>-.314</td>
<td>-.411</td>
<td>-.161</td>
<td>-.430</td>
</tr>
<tr>
<td>S2</td>
<td>-.104</td>
<td>-.063</td>
<td>-.082</td>
<td>-.134</td>
</tr>
<tr>
<td>S3</td>
<td>-.214</td>
<td>-.192</td>
<td>-.152</td>
<td>-.309</td>
</tr>
<tr>
<td>U2</td>
<td>-.036</td>
<td>-.048</td>
<td>-.032</td>
<td>-.071</td>
</tr>
<tr>
<td>U3</td>
<td>-.094</td>
<td>-.144</td>
<td>-.057</td>
<td>-.195</td>
</tr>
<tr>
<td>P2</td>
<td>-.123</td>
<td>-.062</td>
<td>-.086</td>
<td>-.089</td>
</tr>
<tr>
<td>P3</td>
<td>-.386</td>
<td>-.396</td>
<td>-.329</td>
<td>-.261</td>
</tr>
<tr>
<td>A2</td>
<td>-.071</td>
<td>-.068</td>
<td>-.124</td>
<td>-.062</td>
</tr>
<tr>
<td>A3</td>
<td>-.236</td>
<td>-.367</td>
<td>-.325</td>
<td>-.144</td>
</tr>
</tbody>
</table>

Pediatric EQ-5D-Y *

- Recently proposed “child-friendly” version of the ED-5D
  - Children’s preference scores currently unavailable
- Same 5 (renamed) domains:
  - Mobility
  - Looking after myself (Self-care)
  - Doing usual activities (Usual activities)
  - Having pain or discomfort (Pain/discomfort)
  - Feeling worried, sad or unhappy (Anxiety/depression)


Pediatric EQ-5D-Y (2)

- Main difference in question wording occurs in most severe level of each domain
  - “a lot of problems walking about” vs “confined to bed”
  - “a lot of problems washing and dressing” vs “unable to wash or dress”
  - “a lot of problems doing my usual activities” vs “unable to perform my usual activities”
  - “a lot of pain and discomfort” vs extreme pain and discomfort
  - “very worried, sad or unhappy” vs “extremely anxious or depressed”
HUI2

- The HUI2 has 7 domains with varying numbers of levels depending on the domain
- The domains and number of levels include:
  - Sensory with 4 levels
  - Mobility with 5
  - Emotion with 5
  - Cognition with 4
  - Self-care with 4
  - Pain with 5
  - Fertility with 3
- The multiple levels of the seven domains can be used to define 24,000 health states

HUI2 Scoring Rule

- The HUI2 has 2 multiplicative scoring rules that were derived from responses of 293 parents of school children drawn from the general population in Canada
  - Because the rules were initially developed to evaluate a therapy for childhood cancer
- We focus on the utility scoring rule that was developed by use of standard gambles
- At least one other scoring rule has been proposed

Scoring the HUI2

<table>
<thead>
<tr>
<th>Sensation</th>
<th>Level</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Able to see, hear, and speak normally for age</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>2. Requires equipment to see or hear or speak</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>3. Sees, hears, or speaks with limitations even with equipment</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>4. Blind, deaf, or mute</td>
<td>0.61</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mobility</th>
<th>Level</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Able to walk, bend, lift, jump and run normally for age</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>2. Walks, bends, lifts jumps or runs with some limitations</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>3. Requires mechanical equipment</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>4. Requires the help of another person to walk or get around</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>5. Unable to control or use arms and legs</td>
<td>0.58</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Level</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Generally happy and free from worry</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>2. Occasionally fretful, angry, irritable, anxious, depressed</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>3. Often fretful, angry, irritable, anxious, depressed</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>4. Almost always fretful, angry, irritable, anxious, depressed</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>5. Extremely fretful, angry, irritable, or depressed</td>
<td>0.53</td>
<td></td>
</tr>
</tbody>
</table>
Scoring the HUI2

Cognition
1. Learns and remembers normally for age 1.00
2. Learns and remembers more slowly than normal for age 0.95
3. Learns and remembers very slowly 0.88
4. Unable to learn and remember 0.65

Self-Care
1. Eats, bathes, dresses and uses the toilet normally for age 1.00
2. Eats, bathes, dresses or uses the toilet independently but... 0.97
3. Requires mechanical equipment to eat, bathe, dress 0.91
4. Requires the help of another person to eat, bathe, dress 0.80

Scoring the HUI2

Pain
1. Free of pain and discomfort 1.00
2. Occasional pain 0.97
3. Frequent pain. Discomfort relieved by oral medicines 0.85
4. Frequent pain. Discomfort requires prescription narcotics 0.64
5. Severe pain 0.38

Fertility
1. Able to have children with a fertile spouse 1.00
2. Difficulty in having children with a fertile spouse 0.97
3. Unable to have children with a fertile spouse 0.88

Scoring the HUI2

• Scoring formula:
  \[ 1.06 (w_1 \times w_2 \times w_3 \times w_4 \times w_5 \times w_6 \times w_7) - 0.06 \]

<table>
<thead>
<tr>
<th>Domain</th>
<th>Level</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensory</td>
<td>2</td>
<td>0.95</td>
</tr>
<tr>
<td>Mobility</td>
<td>3</td>
<td>0.84</td>
</tr>
<tr>
<td>Emotional</td>
<td>2</td>
<td>0.93</td>
</tr>
<tr>
<td>Cognitive</td>
<td>3</td>
<td>0.88</td>
</tr>
<tr>
<td>Self-care</td>
<td>2</td>
<td>0.97</td>
</tr>
<tr>
<td>Pain</td>
<td>4</td>
<td>0.64</td>
</tr>
<tr>
<td>Fertility</td>
<td>2</td>
<td>0.97</td>
</tr>
<tr>
<td>⋮</td>
<td></td>
<td>0.393</td>
</tr>
</tbody>
</table>

\[ (1.06 \text{ score}) - 0.06 \]

\[ 0.357 \]
HUI3

- The HUI3 has 8 domains each with 5 or 6 levels depending on the domain. The domains and number of levels include:
  - Vision with 6 levels
  - Hearing with 6
  - Speech with 5
  - Ambulation with 6
  - Dexterity with 6
  - Emotion with 5
  - Cognition with 6
  - Pain with 5

- The levels of the domains can be used to define 972,000 health states

HUI3 Scoring Rule

- As with the HUI2, the HUI3 has two multiplicative scoring rules
- For the HUI3, they were derived from responses from a random sample of 256 adults drawn from the general population in Hamilton, Ontario
- Also, as with the HUI2, we focus on the utility scoring rule that was developed by use of standard gambles

<table>
<thead>
<tr>
<th>Scoring the HUI3</th>
<th>Description</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Able to see well enough to read ordinary newsprint and recognize a friend on the other side of the street, without glasses or contact lenses</td>
<td>1.00</td>
</tr>
<tr>
<td>2.</td>
<td>Able to see well enough to read ordinary newsprint and recognize a friend on the other side of the street, but with glasses</td>
<td>0.98</td>
</tr>
<tr>
<td>3.</td>
<td>Able to read ordinary newsprint with or without glasses but unable to recognize a friend on the other side of the street, even with glasses</td>
<td>0.89</td>
</tr>
<tr>
<td>4.</td>
<td>Able to recognize a friend on the other side of the street with or without glasses but unable to read ordinary newsprint, even with glasses</td>
<td>0.84</td>
</tr>
<tr>
<td>5.</td>
<td>Unable to read ordinary newsprint and unable to recognize a friend on the other side of the street, even with glasses</td>
<td>0.75</td>
</tr>
<tr>
<td>6.</td>
<td>Unable to see at all</td>
<td>0.61</td>
</tr>
</tbody>
</table>
Scoring the HUI3

- Scoring formula:
  \[ U^* = 1.371(w_1 \times w_2 \times w_3 \times w_4 \times w_5 \times w_6 \times w_7 \times w_8) - 0.371 \]

<table>
<thead>
<tr>
<th>Domain</th>
<th>Level</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vision</td>
<td>2</td>
<td>0.98</td>
</tr>
<tr>
<td>Hearing</td>
<td>1</td>
<td>1.00</td>
</tr>
<tr>
<td>Speech</td>
<td>2</td>
<td>0.94</td>
</tr>
<tr>
<td>Ambulation</td>
<td>3</td>
<td>0.86</td>
</tr>
<tr>
<td>Dexterity</td>
<td>2</td>
<td>0.95</td>
</tr>
<tr>
<td>Emotion</td>
<td>3</td>
<td>0.85</td>
</tr>
<tr>
<td>Cognition</td>
<td>2</td>
<td>0.92</td>
</tr>
<tr>
<td>Pain</td>
<td>2</td>
<td>0.96</td>
</tr>
</tbody>
</table>

\[(1.371 \text{ score}) - 0.371 = 0.404\]

Comparison of Prescored Instruments

<table>
<thead>
<tr>
<th>EQ-5D</th>
<th>HUI2</th>
<th>HUI3</th>
</tr>
</thead>
<tbody>
<tr>
<td>States with scores ≥ 0.9, N</td>
<td>1</td>
<td>27</td>
</tr>
<tr>
<td>States with scores &lt; 0.0, N</td>
<td>84</td>
<td>63</td>
</tr>
<tr>
<td>Average Score (equal weighting)</td>
<td>0.137</td>
<td>0.286</td>
</tr>
<tr>
<td>Lowest score</td>
<td>-0.594</td>
<td>-0.025</td>
</tr>
</tbody>
</table>
Use of Multiple Instruments in Same Populations

<table>
<thead>
<tr>
<th>N</th>
<th>Instrum</th>
<th>Score</th>
<th>Instrum</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weighted average</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6690</td>
<td>EQ-5D</td>
<td>.700</td>
<td>HUI2</td>
<td>.766</td>
</tr>
<tr>
<td>7570</td>
<td>HUI3</td>
<td>.717</td>
<td>HUI2</td>
<td>.815</td>
</tr>
<tr>
<td>7105</td>
<td>EQ-5D</td>
<td>.711</td>
<td>HUI3</td>
<td>.696</td>
</tr>
<tr>
<td>Simple average</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>EQ-5D</td>
<td>.578</td>
<td>HUI2</td>
<td>.716</td>
</tr>
<tr>
<td>7</td>
<td>HUI3</td>
<td>.577</td>
<td>HUI2</td>
<td>.711</td>
</tr>
<tr>
<td>13</td>
<td>EQ-5D</td>
<td>.686</td>
<td>HUI3</td>
<td>.694</td>
</tr>
</tbody>
</table>

Study samples with significant differences, N *

<table>
<thead>
<tr>
<th>EQ-5D vs HUI2</th>
<th>HUI3 vs HUI2</th>
<th>EQ-5D vs HUI3</th>
</tr>
</thead>
<tbody>
<tr>
<td>3/5</td>
<td>9/7</td>
<td>4/13</td>
</tr>
</tbody>
</table>

Conclusions: Multiple Instruments

- With the exception of 1 large study, there is substantial evidence that the HUI2 yields significantly higher scores than either the EQ-5D or the HUI3.
- Evidence about the EQ-5D and HUI3 is less consistent:
  - Most of the studies, found no significant differences.
  - One large study found that the EQ-5D scores were significantly lower than HUI3 scores.
  - A second large study found that the EQ-5D scores were significantly higher than the HUI3 scores.
- These differences may have to do with the severity of the patients whose preferences were elicited in the two studies.

Relative Responsiveness

- EQ-5D and HUI3:
  - One study found that among patients who improved clinically, the preference scores from the EQ-5D had a significantly larger mean improvement than did the preference scores from the HUI3; among patients who worsened clinically, the preference scores from the EQ-5D had a significantly larger mean decrement than those from the HUI3.
- HUI2 and HUI3:
  - Another study found that the change in HUI3 preferences scores was significantly larger than the change in HUI2 scores.
Superiority?

- Most of the studies that have evaluated correlations between the preference scores have found them to be correlated.
- Most that have evaluated correlations between the preference scores and convergent validity criteria have found them to be correlated.
- Most of the studies that have evaluated responsiveness have concluded that all of the instruments were responsive.
- Most studies have concluded that there is little evidence that one instrument was superior to another.

Some Differences

- Isolated reports of less favorable findings for specific instruments, most often the EQ-5D.
- Luo et al. noted that the EQ-5D demonstrated more ceiling effects than either the HUI2 or HUI3.
- Marra et al. found that the test-retest reliability was acceptable for the HUI2 and HUI3, but was not acceptable for the EQ-5D.
- Suarez-Almazor et al. reported that compared to the HUI2, the EQ-5D did not respond well at 3 months, but did at 6 months.

But Which is Measuring QALYs?

- A number of authors have concluded that while all three instruments appear to be measuring the quality of life, the constructs they are measuring are not identical and their preference scores differ.

"The index scores are not interchangeable in the calculation of longitudinal-based QALYs" (Conner-Spady)
Withholding Judgment

• Given all 3 instruments should be measuring the same construct and that there is little evidence of the superiority of 1 instrument over another, it is problematic that they disagree
• The widespread direct comparison of the instruments does not seem to be providing an answer about which instrument should be used in which circumstances
• Problem partially due to the fact that correlation between the instruments' scores and convergent validity criteria is not sufficient
  – Thus, if 1 instrument has higher correlations with convergent validity criteria, or if it is more responsive than another, that does not necessarily translate into it being a better instrument

Direct Elicitation from Participants

• A second common approach for assessing QALYs in trials is to directly elicit preferences from study participants
• The three most common methods for doing so are:
  – The standard gamble
  – The time trade-off
  – The rating scale
• Sankey's laryngeal cancer example illustrated the use of these method to assess preferences for the duration of morbidity (i.e., the curves he drew)
  – They are used more commonly to develop scores like those derived from the prescored instruments

Standard Gamble

• A standard gamble asks participants to trade-off a certain, intermediate, outcome for a gamble for a better and worse outcome
  – A participant might be asked to choose between living with current health for 10 years versus a p/1-p gamble of living with full function for 10 years or dying immediately
• Standard gambles satisfy the axioms of expected utility theory as proposed by von Neumann and Morgenstern
  – i.e., Preference measured with risk
• In addition to being risk based, standard gambles also require that participants choose between health outcomes
Probability-Equivalent Standard Gamble

- The most common method used to pose the standard gamble is referred to as a probability-equivalent standard gamble.
  - The example above represents such a gamble.
- As in the example, one approach for presenting a probability-equivalent gamble is to describe a level of health for a specific length of survival, for example 10 years or the number of years that represent the participant's life expectancy.
- The participant is asked to identify p such that she is indifferent between the certain outcome and the gamble.

Utility of the Gamble

- We interpret the probability that makes the respondent indifferent between the two choices as the preference or utility score.
  - By indicating indifference, the respondent states that the utility of the certain outcome is identical to the utility of the gamble.
- The utility of the gamble is made up of p times the utility of the best outcome, for example full functioning, which we assume has a utility of 1.0, plus 1 - p times the utility of the worst outcome, which we assume has a utility of 0.

Depiction of Probability-Equivalent Gamble
Time Trade-Off

• A time trade-off asks participants to trade-off morbid years for healthy years
• They are asked to choose between living some length of time – for example, 10 years or the number of years that represent the participant’s life expectancy – with her current health versus living a shorter period of time with fully functional health
• The participant is asked to identify the number of years with fully functional health that makes her indifferent between the longer morbid life expectancy and the shorter fully-functional life expectancy

Time Trade-Offs and Risk

• Unlike standard gambles, time trade-offs do not satisfy the axioms of expected utility theory, because they are not measured with risk
• Like standard gambles, on the other hand, they do require that participants choose between health outcomes

Value of the Time Trade-Off

• To obtain a preference score, we divide the number of healthy years identified by the participant by the number of morbid years
  – For example, if the participant reported that she was indifferent between 7 healthy years and 10 morbid years, the resulting preference score would be 0.7
Rating Scale

- A rating scale – also referred to as visual analog scale or feeling thermometer – asks participants to rate how good or bad their current health is on a 0–1 or 0–100 scale
  - 0 often represents worst imaginable health or death
  - 1 often represents “best imaginable health” or “full health”
- Rating scales can vary in presentation in terms of length of the line, whether they are drawn vertically or horizontally, and whether or not they have intervals marked out with numbers
- Some have argued that having intervals marked out with numbers can induce memory effects and clustering

Rating Scale (II)

- Rating scales neither satisfy the axioms of expected utility theory, nor do they require that participants choose between health outcomes
- If a rating scale ranges between 0 and 1, the point on the line selected by the participant represents her preference score; if the scale ranges between 0 and 100, the point on the line divided by 100 represents this score

Rating Scale Example

To help people say how good or bad a health state is, we have drawn a scale (rather like a thermometer) on which the best state you can imagine is marked by 100, and the worst state you can imagine is marked by 0.

Your own health state today

We would like you to indicate on this scale how good or bad your health is today. Please do this by drawing a line from the box to whichever point on the scale that indicates how good or bad your health is currently.
Use of Multiple Instruments in Same Populations

<table>
<thead>
<tr>
<th></th>
<th>Standard Gamble</th>
<th>Time Tradeoff</th>
<th>Rating Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weighted average</td>
<td>4962</td>
<td>0.872</td>
<td>0.844</td>
</tr>
<tr>
<td>Unweighted</td>
<td>46</td>
<td>0.857</td>
<td>0.825</td>
</tr>
</tbody>
</table>

Conclusions: Multiple Instruments

- Mean scores from the studies from the 46 samples from 27 studies confirm the suggestion in the literature that the difference between mean standard gamble and time tradeoff responses (approximately 0.03) is smaller than the difference between the standard gamble and rating scale (approximately 0.14)
- When individual samples of participants are analyzed, on the other hand, there are large numbers of samples with significant differences (p < 0.05) between all 3 methods
  - The percentage different is higher for standard gambles versus rating scales (84%) than it is for standard gambles versus time trade-offs (62%)

Tengs and Lin Meta-Analyses

- Meta-analyses of responses from patients, caregivers, providers, and members of the community who rated current health or disease scenarios for HIV or stroke
- Time trade-offs appeared to yield the highest preference scores
- Standard gambles were generally 0.1 lower than those for time trade-offs (p=0.16 for HIV and p=0.08 for stroke)
- Rating scale scores were -0.02 less than those for standard gambles when HIV was rated (p = 0.001 compared to time trade-off; NS compared to standard gamble) and -0.11 lower when stroke was rated (p=0.006 compared to time trade-off; p-value not explicitly reported compared to standard gambles)
<table>
<thead>
<tr>
<th>Comparison of Prescored Instruments and Direct Elicitation Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average scores for each instrument and method</td>
</tr>
<tr>
<td>Samples, Participants, Prescored Instrument</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>Weighted average</td>
</tr>
<tr>
<td>EQ-5D</td>
</tr>
<tr>
<td>HUI2</td>
</tr>
<tr>
<td>HUI3</td>
</tr>
<tr>
<td>Simple average</td>
</tr>
<tr>
<td>EQ-5D</td>
</tr>
<tr>
<td>HUI2</td>
</tr>
<tr>
<td>HUI3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comparison of Prescored Instruments and Direct Elicitation Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of studies with statistically significant differences between preference scores derived by direct elicitation and by prescored instruments *</td>
</tr>
<tr>
<td>Samples, Participants, Prescored Instrument</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>EQ-5D</td>
</tr>
<tr>
<td>HUI2</td>
</tr>
<tr>
<td>HUI3</td>
</tr>
</tbody>
</table>

* Statistical tests based on means, SDs, and Ns reported in studies, and do not account for the fact that the same patients responded to all 3 instruments

What to Make of These Findings?

• There is the general recommendation that measures of preference used in economic evaluations should be derived from the general public
  – One rationale for use of pre-scored instruments
• There is some evidence that patients’ preferences for their health are higher than are the general public’s when asked to rate scenarios that mirror the patients’ health, although evidence may not be conclusive
• Appears that rating scale -- which is sometimes considered the least preferred method for the direct elicitation of preferences -- is no worse at reproducing the results of prescored instruments than are the other direct elicitation methods, and may be better
Frequency of Elicitation

- Preferences are usually measured for all study participants at prescheduled intervals, e.g., baseline and semi-annually thereafter.
- Other designs that yield unbiased results include assessment at random intervals or random assignment to assessment intervals.
- One design that will lead to biased results is purposively sampling when a clinical outcome occurs, such as onset of myocardial infarction during follow-up.
Frequency of Elicitation (II)

- The frequency with which preferences should be assessed depends on our beliefs about how rapidly preferences are expected to change, the likely duration of the changes, the length of follow-up, and the resources available for data collection.
- For studies that last several years, we routinely recommend assessing preferences at least twice a year.
  - In a recent long-term clinical trial, we expected an initial rapid change and recommended quarterly assessment during the first year of follow-up.
  - Thereafter we measured preferences semi-annually.

Short-Term "Hellish" Experiences

- Suppose you have a cavity and the dentist plans to drill the tooth for 10 minutes (.000019013 years).
- Suppose you rate these minutes as having a QALY weight of 0 (i.e., you lose 0.000019013 QALYS by having your teeth drilled).
- If an intervention to relieve the pain costs $5, the cost/QALY saved is $263,000 ($5/0.000019013).
- Should we recommend against such interventions?
  - If not, what needs to be changed in our calculation?

Preference Scores for Selected Health States *

<table>
<thead>
<tr>
<th>Health State</th>
<th>Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allergic Rhinitis</td>
<td>0.853</td>
</tr>
<tr>
<td>Psoriasis</td>
<td>0.834</td>
</tr>
<tr>
<td>Migraine</td>
<td>0.806</td>
</tr>
<tr>
<td>Gastritis</td>
<td>0.765</td>
</tr>
<tr>
<td>Depression disorders</td>
<td>0.732</td>
</tr>
<tr>
<td>Blindness / low vision</td>
<td>0.694</td>
</tr>
<tr>
<td>CVA</td>
<td>0.650</td>
</tr>
<tr>
<td>Heart failure</td>
<td>0.636</td>
</tr>
<tr>
<td>Senility w/ psychosis</td>
<td>0.545</td>
</tr>
<tr>
<td>Death</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Toe Nail Fungus Revisited

- Sullivan and Gushchavan report that "Disease of Nail" has a preference score of 0.775.
- If we assume that 0.775 applies to toe nail fungus days, .87 (Luo et al.) applies to the remaining days during the projected 3 years of follow-up, AND that the data in Gupta were otherwise appropriate, we would conclude:

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Total Cost</th>
<th>SFD</th>
<th>QALYs *</th>
<th>C/Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ciclopirox</td>
<td>953.6</td>
<td>563</td>
<td>2.4715</td>
<td></td>
</tr>
<tr>
<td>Itracon pulse</td>
<td>1232.1</td>
<td>612</td>
<td>2.4843</td>
<td>21,797</td>
</tr>
<tr>
<td>Fluconazole</td>
<td>1303.4</td>
<td>620</td>
<td>2.4864</td>
<td>33,810</td>
</tr>
</tbody>
</table>

* ((sfd*.87) + ((1095-sfd)*.775)) / 365; c/q of 9238 and 14,490 if sfd=1

Literature-Based Sources

- Tufts Medical Center, Institute for Clinical Research and Health Policy Studies, Center for the Evaluation of Value and Risk in Health, CEA Registry
  - www.cearegistry.org
  - OR
  - https://research.tufts-nemc.org/cear/

Choice Between Instruments/Methods

- Nothing we have said runs counter to the recommendation that we should measure both the general public’s and patients’ preferences in trials.
- But our review has not led us to strong conclusions about which methods are best for these measurements.
Choice Between Instruments/Methods (II)

- Getting preferences exactly right is complicated, both because human preferences are so variable and have so many determinants and because all the techniques used to measure preferences are flawed
- Many of the “recommendations” from commentators seem to be based on theories that ignore this complexity and these flaws
  – Yes, it would be easy to recommend that you assess the sensitivity of your results to the source of preference scores, but such a strategy is costly
- In sum, it is not clear that strong recommendations about the adoption of specific methods or instruments are supportable

Implementation Issues, Direct Elicitation

Open- Versus Closed-Ended Questions

- Standard gambles and time trade-offs can be administered by use of a single open-ended question,
  – “Which p makes you indifferent?” or “How many years with full function make you indifferent?”
- They are more commonly administered by use of a series of close-ended questions
  – e.g., “Would you rather live with your current health for 10 years or would you choose a gamble in which you have a 90% chance of living 10 fully functional years and a 10% chance of dying immediately.”
  – The probabilities are changed and the question repeated until the respondent reports she is indifferent between the options
Search Procedures

• When offered as a series of close-ended questions, the questions can:
  – Ping pong from high to low to high
  – Offer probabilities or years of healthy survival in steps from the maximum to the minimum (titration down)
  – Offer them from the minimum to the maximum (titration up)
  – Be posed by use of interval division search strategies (bisecting search routines)

Effects of Search Procedures

• Lenert et al. have reported that different search procedures can have strong and persistent effects on reported preference scores for both standard gambles and time trade-offs
  – Supports findings of an earlier study by Percy and Llewellyn-Thomas
• Hammerschmidt et al., on the other hand, did not see significant differences between the results of mailed questionnaire standard gambles that used top-down versus bottom-up search procedures
  – Supports earlier findings by Tsevat et al.

Time Horizon

• Preferences for highly confining health states appear to be a decreasing function of time, whereas preferences for inconvenient health states appear to be an increasing function of time

Torrance et al., 1972
Effects of Different Time Horizons

- Most investigators who have empirically assessed the effect of time horizon have found that the longer the time horizon, the smaller the resulting preference scores.
  - Finding holds for standard gambles, time trade-offs, and rating scales.

<table>
<thead>
<tr>
<th>Morbid Years</th>
<th>Healthy Years</th>
<th>TTO Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>12.5</td>
<td>0.5</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>0.7</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>1.0</td>
</tr>
</tbody>
</table>

McNeil et al. 1981

What Time Horizons Have Investigators Used?

- Out of 27 studies that asked patients to rate their current health by use of the standard gamble, time trade-off, and rating scale:
  - 7 used time horizons of 10 years or less
  - 7 used horizons of 20–30 years
  - 12 used life expectancy as the time horizon
  - (3 did not identify the time horizon)
- It is unclear how much variability of results in the literature arises because of the use of these different time horizons.

Methods of Administration

- Standard gambles and time trade-offs are most commonly administered by use of an interview, often with the aid of chance boards, decision wheels, and pie charts.
- Rating scales are usually self-administered as part of an interview.
- All three methods can be self-completed by participants.
Computer-Based Administration

- There are a number of computer-directed questionnaires that can be used for their administration:
  - U-Titer
  - U-Maker
  - iMPACT
  - The Functional Limitation and Independence Rating (FLAIR) has been developed to support preference assessment in older, computer-inexperienced participants
- Investigators have also designed internet-based surveys of preference for specific studies

Telephone Surveys

- van Wijck et al. have reported that telephone interviews (preceded by a mailed survey) yield standard gamble and time trade-off results that are similar to those obtained by face-to-face interview

Mailed Surveys

- There is good evidence of the feasibility of use of rating scales in mailed, self-completed surveys [63,64]
- Evidence for the feasibility of mailed, self-completed standard gambles appears more mixed
  - Ross et al. and Littenberg et al. have reported that a one-page paper standard gamble is a reliable measure of patient preference and is suitable for use in mailed surveys
  - Hammerschmidt et al., on the other hand, have reported substantial feasibility problems for mailed, self-completed standard gambles
General Practicality

- Green et al. report that there is substantial evidence supporting all 3 methods' practicality in terms of completion and response
  - They discount claims that standard gambles are too complex or not intuitively obvious to participants
  - But they do note that rating scales may be "slightly better in terms of response rate and cost"
- They also note that standard gambles and time trade-off methods may "result in a larger number of refusals, missing values, and inconsistent responses" than do rating scales

General Practicality (II)

- Woloshin et al. have more recently raised concerns about the quality of the results from standard gambles and time trade-offs among less numerate participants
- Green et al. report that all the three methods have acceptable levels of reliability, although they found some evidence that the time trade-off may have slightly better test–retest performance