Cost-Utility Analysis (CUA), part II

Marcelo Coca Perraillon
University of Colorado
Anschutz Medical Campus

Cost-Effectiveness Analysis
HSMP 6609
2016
Last class we saw the way health states are valued in cost-effectiveness studies.

The objective is to come up with a number that represents “quality” of life, which is measured in terms of preferences over health states.

Scales like the SF-36, SF-12, or even the EQ-5D questionnaire, measure health status but do not value them.

In practice, studies use different approaches: ask participants to value states, ask family to value states, ask doctors to value states, use valuations from representative samples.

The Panel on CE recommended valuation of health states by a representative sample in the reference case (see Chapter 4).
1. For a Reference Case analysis, incorporation of morbidity and mortality consequences into a single measure should be accomplished using QALYs.

2. In general, since lives saved or extended by an intervention will not be in perfect health, a saved life year will count as less than 1 full QALY.

3. To satisfy the QALY concept, the quality weights must be preference-based, interval-scaled, and measured or transformed onto an interval scale where the reference point “death” has a score of 0.0 and the reference point “optimal health” has a scale of 1.0.

4. Community preferences for health states are the most appropriate ones for use in a Reference Case analysis.

5. When community preferences are used and the program (treatment or prevention) is related to an illness or condition, a sensitivity analysis that furnishes information on the preferences of persons with the condition will provide important ancillary information.

6. If distinct subgroup preferences are identified that will markedly affect a C/E ratio, a Reference Case analysis should provide this information and conduct separate sensitivity analyses that reflect this difference.

7. The health-related quality of life of those whose lives have been saved or extended by a health intervention may be influenced by characteristics such as the age, gender, race, or socioeconomic status of the population involved. This may affect a Reference Case analysis in ways that are ethically problematic. In these instances, we recommend that sensitivity analysis be conducted to indicate explicitly how the analysis is affected by these characteristics.
Outline for today

- Other instruments to measure preferences (there are many options)
- Mapping (prediction) models using general functioning scales
- General recommendations
- The concept of a \( \frac{\$}{QALY} \) threshold
Measuring preferences

- Although the EQ-5D is widely used, it’s not the only way to measure preferences.
- Other common instruments include the Health Utility Index (HUI), Quality of Well-Being (QWB), SF-6D, and HALex.
- The basic principles we covered last class when talking about the EQ-5D apply to other instruments:
  1. Define health states using the instrument.
  2. Elicit preferences for the health states (community or patients in intervention).
- In the US, not all instruments have preference scores representative of the US population.
- In the past, many US studies used preferences elicited in the UK or Canada.
The Beaver Dam study

The Beaver Dam Health Outcomes study (named after a town in Wisconsin) is one of the oldest projects designed to measure health status and health-related quality of life. Started in 1990 with three objectives:

1. Provide a catalog of health-state utilities
2. Obtain longitudinal estimates of health-status transition probabilities
3. Compare health status instruments

They used the Quality of Well-being (QWB) and the SF-36 to measure health status but elicited preferences using the QWB instrument.
The QWA measures health status in terms of mobility, physical activity, social activity, and symptoms.

The original QWA was too long; there is a shorter version: QWB-SA.

The Beaver Dam study elicited preferences for health states defined by the QWB, not the SF-36.

This is a problem because the SF-36 was widely used at the time. Why did they do that?
The problem of common instruments that measure health status

- They define **too many** health states
- The SF-36 has 36 questions, some of them with 10 possible answers. With only 2 possible answers per question: $2^{36} = 68,719,476,736$ health states. You see the problem
- That’s the main reason instruments designed to measure health states to elicit preferences must define few states
- As part of the Beaver Dam study, they used statistical models to predict QWB-derived preferences from SF-36 scores (Fryback, 1997)
- Their models did just fine and allowed researchers to conduct CUAs from studies that asked respondents to complete the SF-36
Whose preferences?

- As we saw with the EQ-5D, and according to the recommendation of the Panel on CE, we would like to elicit preferences representative of the US population.
- The with the Beaver Dam study is that the preferences they elicited were representative of Beaver Dam, WI (99.3% percent white).
- Not exactly the ideal, but better than nothing.
- No EQ-5D or even SF-12 at the time.
Other options

- Your textbook describes other instruments
- For example, the SF-6D was designed to convert health profiles from the SF-36 into preferences
- Problem is, there is no study that has done the same for a sample representative of the US population
- An exception is Craig et al (2013), who derived preferences for the SF-6D in the US
- But their sample is not representative of the US
- Same with the QWB, which has preference scores for a sample of San Diego, CA
## Health Utilities Index (HUI2)

### Table 5.4 Health Utilities Index mark 2 classification system

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Level</th>
<th>Level description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensation</td>
<td>1</td>
<td>Ability to see, hear, and speak normally for age</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Requires equipment to see or hear or speak</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Sees, hears, or speaks with limitations even with equipment</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Blind, deaf, or mute</td>
</tr>
<tr>
<td>Mobility</td>
<td>1</td>
<td>Able to walk, bend, lift, jump, and run normally for age</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Walks, bends, lifts, jumps, or runs with some limitations but does not require help</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Requires mechanical equipment (such as cane, crutches, braces, or wheelchair) to walk or get around independently</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Requires the help of another person to walk or get around and requires mechanical equipment as well</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Unable to control or use arms and legs</td>
</tr>
<tr>
<td>Emotion</td>
<td>1</td>
<td>Generally happy and free from worry</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Occasionally fretful, angry, irritable, anxious, depressed, or suffering ‘night terrors’</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Often fretful, angry, irritable, anxious, depressed, or suffering ‘night terrors’</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Almost always fretful, angry, irritable, anxious, depressed</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Extremely fretful, angry, irritable, anxious, or depressed; usually requiring hospitalization or psychiatric institutional care</td>
</tr>
<tr>
<td>Attribute</td>
<td>Level</td>
<td>Level description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------</td>
<td>------------------------------------------------------------</td>
</tr>
<tr>
<td>Cognition</td>
<td>1</td>
<td>Learns and remembers schoolwork normally for age</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Learns and remembers schoolwork more slowly than classmates judged by parents and/or teachers</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Learns and remembers very slowly and usually requires special educational assistance</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Unable to learn and remember</td>
</tr>
<tr>
<td>Self-care</td>
<td>1</td>
<td>Eats, bathes, dresses, and uses the toilet normally for age</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Eats, bathes, dresses, or uses the toilet independently with difficulty</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Requires mechanical equipment to eat, bathe, dress, or use the toilet independently</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Requires the help of another person to eat, bathe, dress, or use the toilet</td>
</tr>
<tr>
<td>Pain</td>
<td>1</td>
<td>Free of pain and discomfort</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Occasional pain. Discomfort relieved by non-prescription drugs or self-control activity without disruption of normal activities</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Frequent pain. Discomfort relieved by oral medicines with occasional disruption of normal activities</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>Frequent pain, frequent disruption of normal activities. Discomfort requires prescription narcotics for relief</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>Severe pain. Pain not relieved by drugs and constantly disrupts normal activities</td>
</tr>
<tr>
<td>Fertilitya</td>
<td>1</td>
<td>Able to have children with a fertile spouse</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>Difficulty in having children with a fertile spouse</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>Unable to have children with a fertile spouse</td>
</tr>
</tbody>
</table>
One option that has become popular is to use statistical models to predict the EQ-5D preference index (the -1 to 1 number) using other instruments, particularly the SF-12.

The SF-12 is a shorter version of the SF-36.

You are probably noticing a trend: the objective is to shorten these instruments so they are easier to administer, regardless of whether they define few states or not.

The latest in this area is using adaptive testing derived from Item Response Theory (IRT) (like the GRE).
(6) In general, would you say your health today is:
   1. Excellent
   2. Very good
   3. Good
   4. Fair
   5. Poor

The following two questions are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?

(7) Moderate activities such as moving a table, pushing vacuum cleaner, bowling, or playing golf:
   1. No. Not limited at all
   2. Yes. Limited a little
   3. Yes. Limited a lot

(8) Climbing several flights of stairs:
   1. No. Not limited at all
   2. Yes. Limited a little
   3. Yes. Limited a lot

During the past 4 weeks, have you had any of the following problems with your work or regular activities as a result of your physical health?

(9) Accomplished less than you would like?
   1. No
   2. Yes
The reason for the proliferation of mapping methods in the US is that the Medical Expenditure Panel Survey (MEPS) in 2001-2003 asked 15,000 respondents (per year!) to complete both the EQ-5D and the SF-12.

With so much data, it is possible to evaluate the prediction performance of statistical models.

As it turns out, though, it is not exactly an easy problem.
Distribution of the EQ-5D in the US (2001 MEPS data)
Distribution of the EQ-5D by age and medical condition

Figure 1  Distribution of EQ-5D by age group and medical condition. Data source: MEPS, 2000. All sample (A); by age group (B–E), and for selected self-reported conditions (F–I). “Any condition” refers to those who have heart disease, stroke, and/or diabetes. Some individuals have more than 1 condition.
Why hard to predict?

- The EQ-5D preference index doesn’t distribute normal
- It’s bounded at 1 and -0.54
- Has three distinct modes
- On the other hand, it does correlate well with SF-12 summary scores. Correlation with SF-12 summary scores is over 0.6
- Not surprising: people prefer good health to bad health, especially when the health state is hypothetical
Researchers tried different alternatives: linear regression, two-part models, Tobit models, multinomial models (to predict responses)

Predictors are summary scores of the SF-12 (mental and physical), age, sex, education (and many more sometimes)

Problem is predictions are systematically biased

1. Over- and under-predicts the upper and lower values
2. Shrinks variance (regression to the mean)
Alternative: assume a mixture distribution

- Mixtures models assume that the distribution is mixture, with probability, of two or more distributions
- They can be used to classify observation into classes
- Example, height distribution in the population
NHANES 2009-2010 data on height (combined and by sex, pink—women)
For the EQ-5D

\[ f(y)dy = \left[ \sum_{j=1}^{2} \pi_j \frac{1}{\sigma_j} \phi \left( \frac{y-x'_j\beta_j}{\sigma_j} \right) dy \right]^{(1-d)} \left[ \pi_0 + \sum_{j=1}^{2} \pi_j \Phi \left( -\frac{x'_j\beta_j}{\sigma_j} \right) \right]^d \]

For more info, see paper:
http://mdm.sagepub.com/content/35/7/888.full.pdf+html
Observed

A) All (n=14,241)

Linear Model (OLS)

Two-part Model

Mixture 1 (WA)

Mixture 2 (WA)

Mixture 2 (CEC)

Mixture 2 (CEC) - Any Condition

Stroke, diabetes or heart disease
So which one should you use?

- Hard to come up with a good answer. It does matter which one you use and it makes a difference, but how much of a difference depends on the situation

- General advice:
  1. If you can design the study, use the EQ-5D and other instruments like the SF-12
  2. If the EQ-5D is not specific enough for the condition you’re studying, search the literature for other instruments that have been used for the same condition (disease specific)
  3. If there are instruments that have been mapped to preferences, consider those. For example, SF-12 to EQ-5D or FACT-G to EQ-5D...
  4. In many situations, you don’t design the study. This is a common problem in the US, especially from studies that use clinical trial data. Few options other than mapping

- There is a paper published about this topic in every single issue of MDM
The EQ-5D is not the only option, but in the US, not a lot of options if you want to get preferences that are representative of the US population.

Remember that your textbook is from the UK; this problem is not mentioned.

You may wonder, why bother with QALYs at all? Why not natural units or relevant scales that are disease specific?

All together now: **Comparability**!

If we want to use CEA to allocate resources, we need to be able to compare studies.

Which is a good introduction to the idea of **thresholds**.
Thresholds

- I have mentioned thresholds several times
- The idea is that there is a number that can be used as a guide to decide if an intervention is cost effective or not
- We saw that if \( ICER = \frac{\Delta C}{\Delta QALY} < R_T \), the intervention is cost-effective
- We also saw (with some algebra) that you can think of this number as both, willingness to pay for a QALY or the opportunity cost of the intervention
- In the US, for many years, that number was $50,000 per QALY, and lately, $100,000 per QALY
- **Where does that number come from?**
In short, the answer appears to be that the number came from thin air.

Rumor has it that it came from Medicare’s decision to approve end-stage renal disease (ESRD) coverage in the 1970s.

At that time, the ICER for dialysis was about 50K per QALY.

In other words, the government, by deciding that 50K per QALY was “worth it,” provided a sort of endorsement.

Turns out that the ICER for dialysis was about 25K and the approval of ESRD had little to do with CEA.

Neumann et al (2014) call the link between the 50K and ESRD “an urban legend”.
So where did it come from?

- Grosse (2008) tracks the history of the 50K threshold

<table>
<thead>
<tr>
<th>Year</th>
<th>Action</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1992</td>
<td>First widespread use of CE thresholds, $20,000 and $100,000 per QALY, in Canada</td>
<td>[33]</td>
</tr>
<tr>
<td>1992</td>
<td>First publication to use a CE threshold of $50,000 per QALY</td>
<td>[87]</td>
</tr>
<tr>
<td>1995</td>
<td>Second publication to use CE threshold of $50,000 per QALY and first to promote it</td>
<td>[92]</td>
</tr>
<tr>
<td>1995</td>
<td>Widely cited article using CE threshold of $50,000 per LY published</td>
<td>[61]</td>
</tr>
<tr>
<td>1996</td>
<td>Panel on Cost-Effectiveness in Health and Medicine issues report that mentions $50,000 per QALY figure and discourages its use as a criterion for CE</td>
<td>[127,128]</td>
</tr>
<tr>
<td>1996</td>
<td>Use of $50,000 per QALY or LY as a common point of comparison begins to take off</td>
<td>[70,76,95]</td>
</tr>
<tr>
<td>1997–1998</td>
<td>Experts begin referring to anything costing less than $50,000 per QALY as ‘economically attractive’ or ‘reasonably efficient’</td>
<td>[73,99]</td>
</tr>
</tbody>
</table>

The $50K per QALY is an arbitrary decision rule that lacks theoretical or empirical justification and is in any case outdated.

At the very least, *adjust for inflation*: $50,000 \times (1.03)^{25} = $104,689

Medical inflation is a lot higher: $50,000 \times (1.045)^{25} = $150,272

The $50K per QALY might have had something to do with estimates for hemodialysis for ESRD in the early 1990s, in terms of cost per life-year (LY) saved rather than cost per QALY (Grosse, 2008).

The $50K is not the only threshold.
Other thresholds

- From Neumann et al (2014)

| Cost-Effectiveness Thresholds Referred to by Authors of U.S.-Based Cost-Utility Analyses, 1990–2012.* |
|-------------------------------------------------|------------------|------------------|------------------|
| $50,000 per QALY                             | 19.3 percent      | 36.6 percent      | 36.9 percent      |
| $100,000 per QALY                            | 6.3 percent       | 7.8 percent       | 16.9 percent      |
| Both $50,000 and $100,000 per QALY            | 3.9 percent       | 19.9 percent      | 23.7 percent      |
| Other                                        | 18.4 percent      | 10.6 percent      | 7.4 percent       |
| No threshold referenced                      | 51.9 percent      | 25.1 percent      | 15.3 percent      |

* Data are from the Tufts Medical Center Cost-Effectiveness Analysis Registry (www.cearegistry.org). QALY denotes quality-adjusted life-year.
How could we come up with a threshold?

- Remember that we could maximize benefits (years of life or QALY) by arranging interventions from lowest to higher ICER and allocating a budget starting with the lowest ICER.
- The ICER of the last intervention (the highest ICER added to the budget) covered is the maximum society is willing to pay for health gains.
- The problem is that we don’t have that kind of information. We have seen that studies don’t use the same methods and we don’t have ICERs for a lot of interventions, technologies, or drugs. Besides, in the US, we just keep expanding the budget...
- In the UK, NICE often said that they didn’t have a hard threshold, but researchers figured that interventions with $ICER < £20K$ were always approved, and those with $ICER < £30$ were approved with some restrictions.
- Now ICER endorses the 20K to 30K range.
How could we come up with a threshold?

- Another way would be to use **revealed preferences**. Based on actual behavior in situations of risk, can we infer what value people place on life?
- I’ll talk more about this next class because it has to do with cost-benefit analysis.
- But the idea is that we could estimate the value of life and the value of one year of life using this method (and then adjust for quality).
- Using this method, the current estimate for the US is a threshold of $110,000 to $160,000 per QALY.
How could we come up with a threshold?

- Another way: what about if we compared increases in health spending over time with increases in health gains to figure out the value of that money in terms of health?
- That method produces a threshold of $200,000 to $300,000 per QALY (Hirth et al, 2000)
- In the UK, a recent study using this method found a value of about £12K to £30K per QALY (Claxton et al, 2015)
So where are we?

■ The only agreement so far is that $50,000 per QALY is too low
■ Other than that, not much agreement on an exact number but agreement on a broad range
■ On the other hand, it makes sense that there shouldn’t be a hard value because there are many ways to come up with a number that is reasonable but has many flaws
■ Neumann et al (2014) recommend using a range of values $50,000, $100,000, and $200,000 per QALY
■ It is also useful to compare the ICERs of similar interventions to have an idea of how a new treatment or drug compares to similar treatments
■ But keep in mind that ICER is always a comparison, so you want to make the same comparison, which currently is $/QALY
Different organizations have different standards. For example, for drugs (Neumann and Cohen, 2015)

<table>
<thead>
<tr>
<th>Institute for Clinical and Economic Review (ICER)</th>
<th>Incremental cost-effectiveness plus care value components</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Comparative clinical effectiveness</td>
</tr>
<tr>
<td></td>
<td>Other benefits and disadvantages</td>
</tr>
<tr>
<td></td>
<td>Contextual considerations</td>
</tr>
<tr>
<td></td>
<td>Budget impact</td>
</tr>
</tbody>
</table>

Cost-effectiveness ratio must not exceed a threshold ranging from $100,000 to $150,000 per QALY. Selection of final threshold is based on: (a) comparative clinical effectiveness, reflecting “judgments of the health benefit magnitude” and “strength of a body of evidence”; (b) other benefits and disadvantages, including such outcomes as factors influencing adherence or return to work; and (c) contextual considerations, including “ethical, legal, or other issues” (e.g., high burden of illness, availability of alternative treatments). Budget impact is acceptable if a drug’s introduction is compatible with an annual health care budget increase of GDP growth plus 1%. ICER reverse-engineers a “value-based price benchmark” that independently satisfies both the cost-effectiveness and budget-impact criteria (see text).
### Other standards

<table>
<thead>
<tr>
<th>American College of Cardiology–American Heart Association (ACC–AHA)</th>
<th>Clinical benefit vs. risks</th>
<th>Magnitude of treatment effect ranges from class I (“benefit [greatly exceeds] risk,” “procedure or treatment is useful or effective”) to class III (“no benefit, or harm,” “procedure or treatment is not useful or effective and may be harmful”). Precision of treatment effect ranges from level A (“data derived from multiple randomized trials or meta-analyses”) to level C (“only consensus opinion of experts, case studies, or standard of care”). Value corresponds to cost-effectiveness thresholds (high: less than $50,000 per QALY; intermediate: $50,000 to $100,000 per QALY; low: more than $150,000 per QALY). The framework lists the clinical benefit and value designations without combining them.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magnitude of net benefit</td>
<td>Precision of estimate based on quality of evidence</td>
<td>Value (cost-effectiveness)</td>
</tr>
</tbody>
</table>
The EQ-5D is not the only instrument to define health states but not a lot of options in the US

Valuation studies are expensive and difficult to do

Valuation studies of instruments that define too many health states are not practical

When designing a study, include instruments that can be translated into preferences (EQ-5D, SF-6D, SF-12)

There is no hard threshold, and there shouldn’t be one, but there is a range of reasonable values