Cost-Utility Analysis (CUA), part II

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Cost-Effectiveness Analysis
HSMP 6609
2020

Outline for today

- Other instruments to measure health states and preferences (there are many options)
- (Very) brief overview of prediction methods
- Review of decision rules using a threshold
- A preview of measuring the value of life
- The $\frac{\$}{QALY}$ threshold to decide on cost-effectiveness and where it came from
- How is the ICER used in practice?

Big picture

- Last class we saw how **health states** are valued in cost-effectiveness studies (**preference elicitation**)
- The objective was to come up with a number that represents "quality" of life, which is measured in terms of **preferences** over health states
- Once we have that number, we can **adjust life years for quality**
- Remember that the number is (for convenience) set to be between 0 and 1, but in the US because of the way the Shaw study was conducted the number can be negative

Big picture

- Also, keep in mind why we talked about **health measurement**
- We talked about it because we do measure outcomes when we study the impact of new technologies/interventions
- We also talked about it because **we wanted to define health states**
- Scales like the SF-36, SF-12, or even the EQ-5D questionnaire, measure *health status* but do not directly measure preferences
- We saw how the time trade-off could be used to translate the EQ-5D into preferences
- 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
- We saw that the recommendation is to use community valuations like the Shaw study in the US (as in the homework)

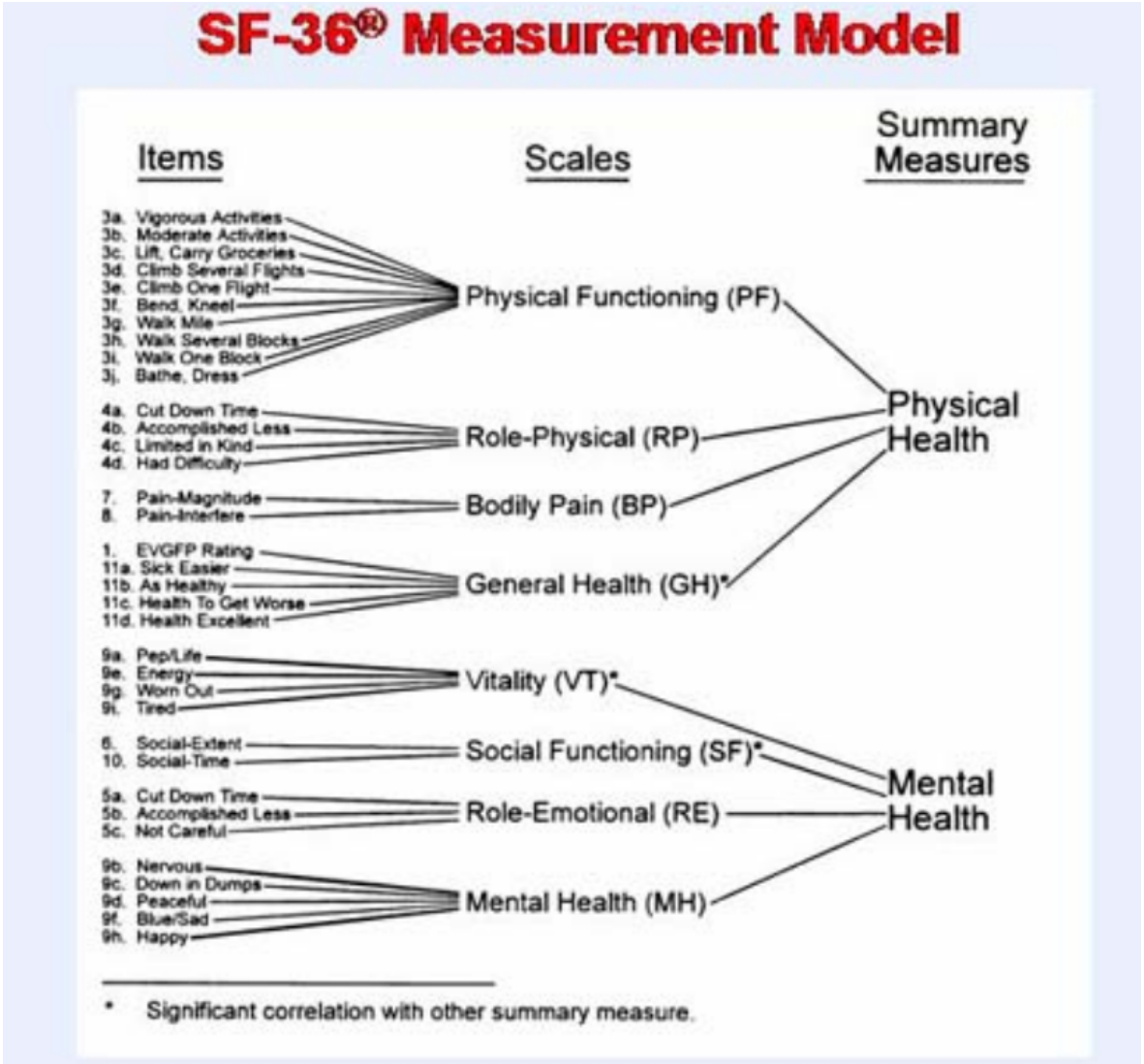
Measuring preferences

- 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 (usually using the time trade-off method)
- In the US, not all instruments have preference scores representative of the US population; it is expensive and time consuming
- Next, a **brief** history of preference elicitation in the US

The SF-36

- The SF-36 is a generic instrument (as opposed to disease specific)
- It measures several domains of health in two categories: physical and mental functioning
- As the name indicates, it has 36 questions
- A lot of work and research goes into the development of these scales
- In fact, there is a whole field dedicated to methods for developing scales: **psychometrics**

SF-36


SF-36® Measurement Model
Items
Scales
Summary Measures
3a. Vigorous Activities
3b. Moderate Activities
3c. Lift, Carry Groceries
3d. Climb Several Flights
3e. Climb One Flight
3f. Bend, Kneel
3g. Walk Mile
3h. Walk Several Blocks
3i. Walk One Block
3j. Bathe, Dress
Physical Functioning (PF)
4a. Cut Down Time
4b. Accomplished Less
4c. Limited in Kind
4d. Had Difficulty
Role-Physical (RP)
7. Pain-Magnitude
8. Pain-Interfere
Bodily Pain (BP)
1. EVGFP Rating
11a. Sick Easier
11b. As Healthy
11c. Health To Get Worse
11d. Health Excellent
General Health (GH)*
Physical Health
9a. Pep/Life
9e. Energy
9g. Worn Out
9i. Tired
Vitality (VT)*
6. Social-Extent
10. Social-Time
Social Functioning (SF)*
5a. Cut Down Time
5b. Accomplished Less
5c. Not Careful
Role-Emotional (RE)
9b. Nervous
9c. Down in Dumps
9d. Peaceful
9f. Blue/Sad
9h. Happy
Mental Health (MH)
Mental Health
* Significant correlation with other summary measure.

Problem with common instruments in CEA

- The problem with common instruments for CEA is that 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
- To elicit preferences, instruments need to define fewer health states
- (**Don't get lost**: not a general problem, a problem for eliciting preferences)
- 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

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 really representative of the US
- Same with the QWB, which now has preference scores for a sample of San Diego, CA

Health Utilities Index (HUI2)

Table 5.4 Health Utilities Index mark 2 classification system

Attribute	Level	Level description
Sensation	1	Ability to see, hear, and speak normally for age
	2	Requires equipment to see or hear or speak
	3	Sees, hears, or speaks with limitations even with equipment
	4	Blind, deaf, or mute
Mobility	1	Able to walk, bend, lift, jump, and run normally for age
	2	Walks, bends, lifts, jumps, or runs with some limitations but does not require help
	3	Requires mechanical equipment (such as cane, crutches, braces, or wheelchair) to walk or get around independently
	4	Requires the help of another person to walk or get around and requires mechanical equipment as well
	5	Unable to control or use arms and legs
Emotion	1	Generally happy and free from worry
	2	Occasionally fretful, angry, irritable, anxious, depressed, or suffering 'night terrors'
	3	Often fretful, angry, irritable, anxious, depressed, or suffering 'night terrors'
	4	Almost always fretful, angry, irritable, anxious, depressed
	5	Extremely fretful, angry, irritable, anxious, or depressed; usually requiring hospitalization or psychiatric institutional care

Health Utilities Index (HUI2)

Attribute	Level	Level description
Cognition	1	Learns and remembers schoolwork normally for age
	2	Learns and remembers schoolwork more slowly than classmates as judged by parents and/or teachers
	3	Learns and remembers very slowly and usually requires special educational assistance
	4	Unable to learn and remember
Self-care	1	Eats, bathes, dresses, and uses the toilet normally for age
	2	Eats, bathes, dresses, or uses the toilet independently with difficulty
	3	Requires mechanical equipment to eat, bathe, dress, or use the toilet independently
	4	Requires the help of another person to eat, bathe, dress, or use the toilet
Pain	1	Free of pain and discomfort
	2	Occasional pain. Discomfort relieved by non-prescription drugs or self-control activity without disruption of normal activities
	3	Frequent pain. Discomfort relieved by oral medicines with occasional disruption of normal activities
	4	Frequent pain, frequent disruption of normal activities. Discomfort requires prescription narcotics for relief
	5	Severe pain. Pain not relieved by drugs and constantly disrupts normal activities
Fertility[a]	1	Able to have children with a fertile spouse
	2	Difficulty in having children with a fertile spouse
	3	Unable to have children with a fertile spouse

Prediction (**mapping**) example

- One option that has become more common is to use statistical models to predict the EQ-5D preference index (the -1 to 1 number) using other instruments
- Simple idea: for example, ask people to complete both instruments (i.e. the EQ-5D and SF-12)then use statistical models to predict the preference index using the SF-12 responses (or summary scores)
- The **reason this works is because both instruments are trying to get at the same idea** (**construct**): how healthy is a person?
- The preference index is a translation: how much do we value being healthy (or no)?

Department of Easy Things to Miss

- We spent a lot of class time talking about eliciting preferences, utility under uncertainty, etc
- We wanted to measure preferences over health states. I said that other instruments measure health status instead of preferences
- **The reason mapping/predicting is not a bad idea is because health status is highly correlated with preferences** over hypothetical health status

SF-12

SF-12

(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 muc

(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

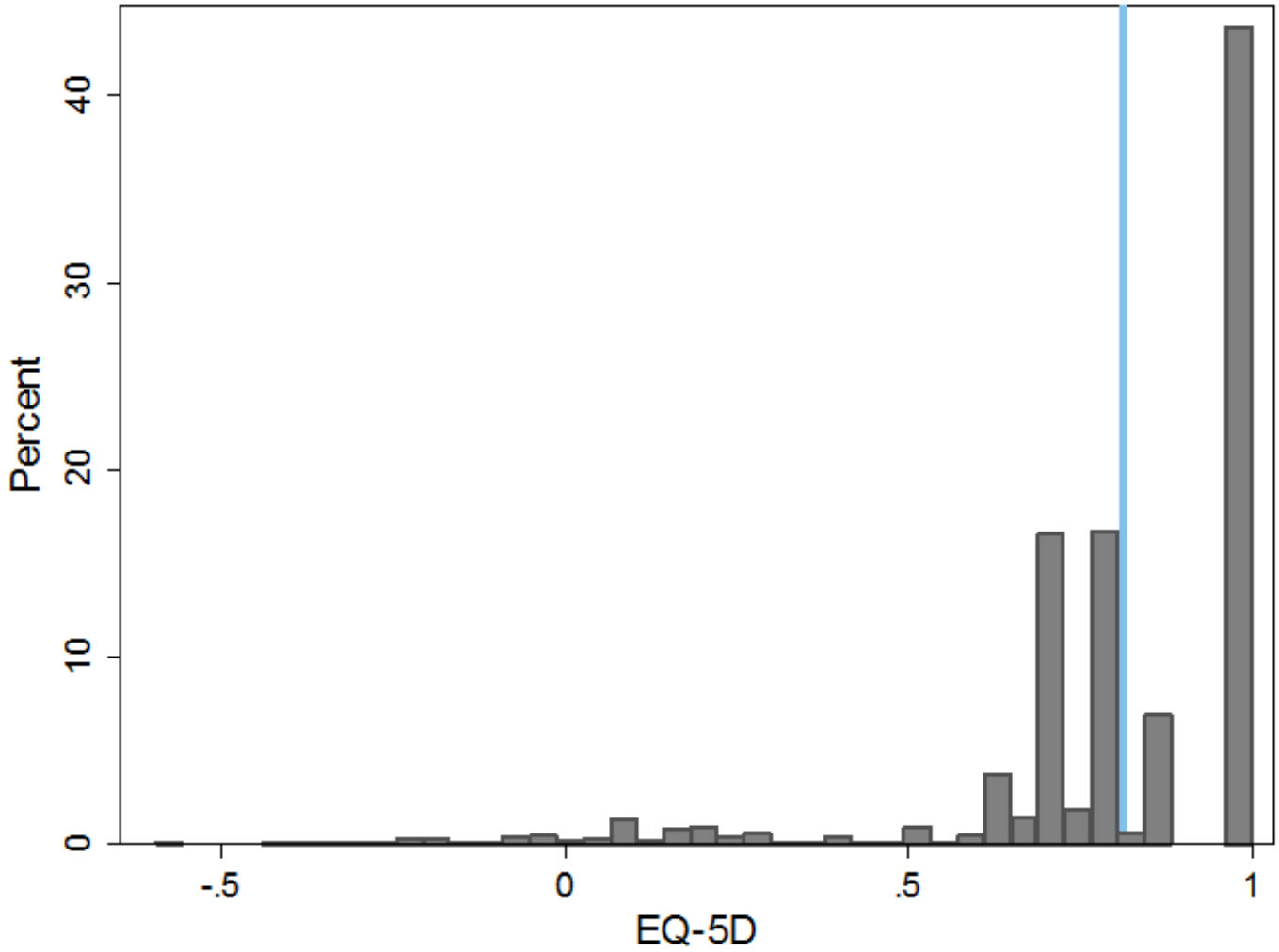
MEPS data

- 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
- By the way, the MEPS has great data on health care utilization, health expenditures, co-sharing, etc
- Free to download `https://www.meps.ahrq.gov/mepsweb/`
- With so much data, it is possible to evaluate the prediction performance of statistical models
- But...

Hard to predict because

- The EQ-5D preference index doesn't distribute normal (or any other standard distribution)
- It's bounded at 1 and -0.54
- Has three distinct modes
- To make a long story short, researchers have proposed many methods, from linear models to finite mixtures to multinomial models
- See my paper in optional readings. I tried a mixture model

Distribtuion of the EQ-5D in the US (2001 MEPS data)



Distribtuion of the EQ-5D by age and medical condition

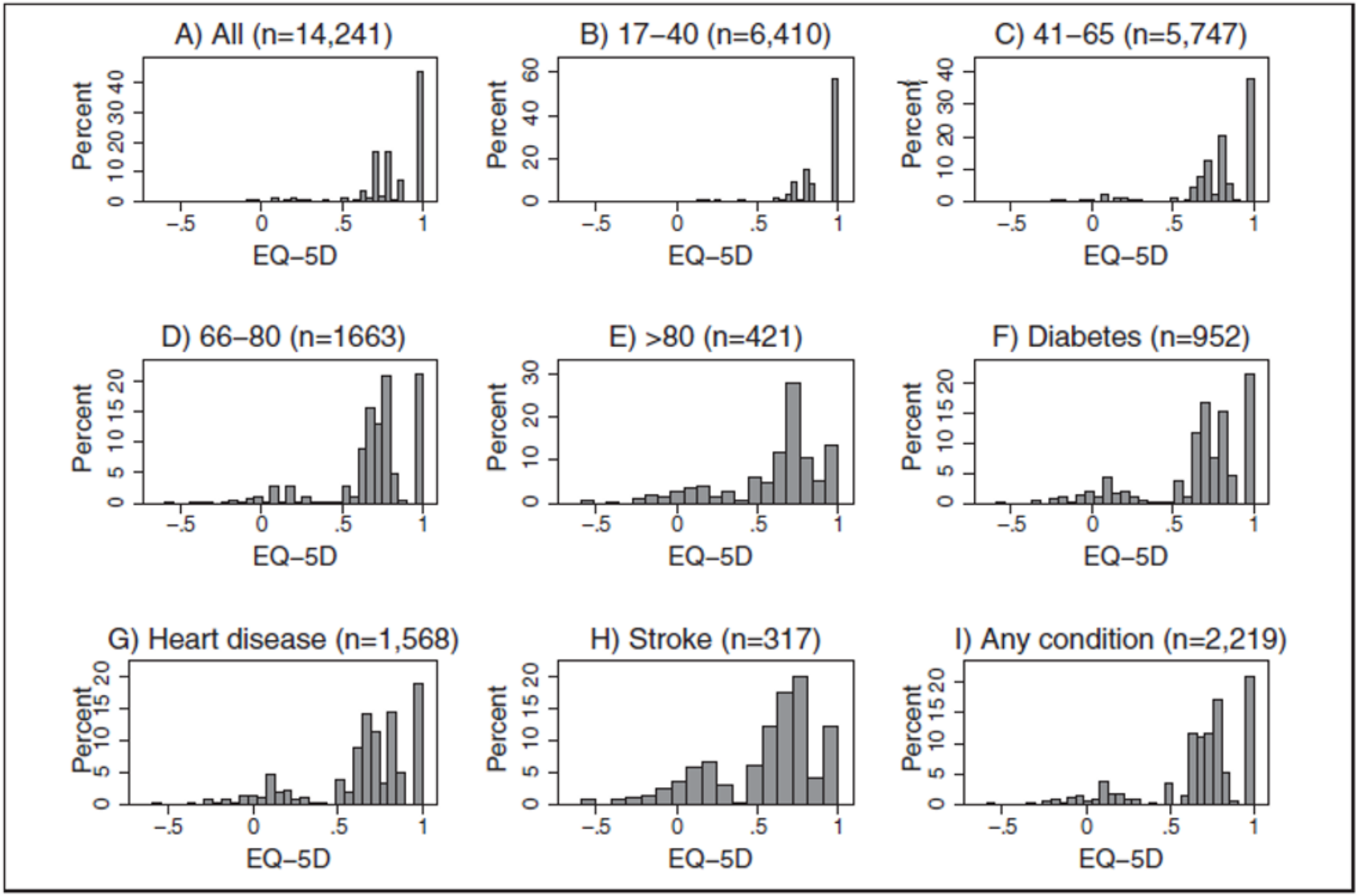


Figure 1 Distribution of EQ-5D-3L by age group and medical condition. Data source: MEPS, 2000. All sample (A); by age group (B–E), and for selected self-reported conditions (F–I). "Any condition" refers to those who have heart disease, stroke, and/or diabetes. Some individuals have more than 1 condition.

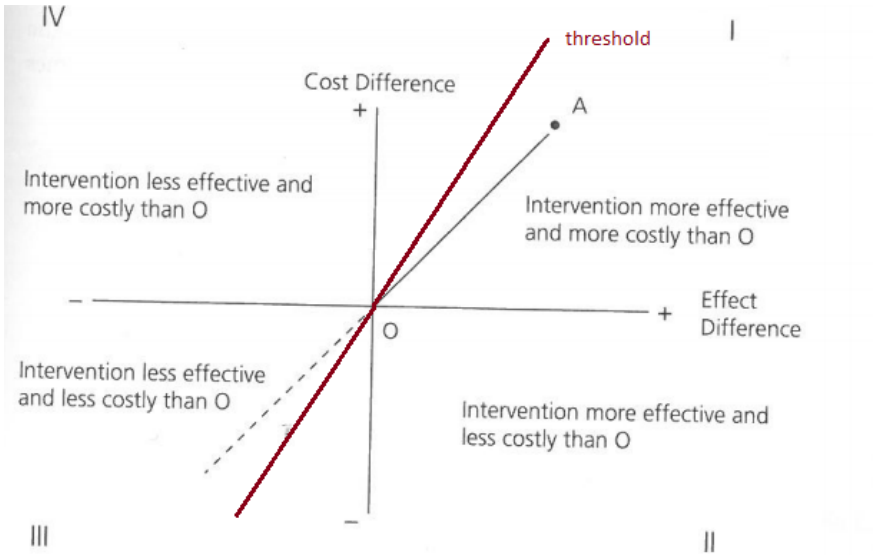
Which one should you use?

- With so many instruments and methods, which one should you use if you need the 0 to 1 preference index?
- 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 common 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. In that case, few options other than mapping
- There seems to be a prediction paper every other month in the literature...

Summary

- 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 that extreme in the UK but the book mentions predictions/mapping
- 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**

The cost-effectiveness plane, again



- Threshold line (R_T): the max the decision maker is willing to pay for a unit of effect
- The most common: \$50K per QALY; now about \$100k to \$150K per QALY in the US
- **Decision rule**: if $ICER_{A,O} < R_T$ or $\Delta C/\Delta E < R_T$

The ICER has some problems

- Some issues with the ICER, some of them not that problematic:
- 1) The ratio of two negative numbers is positive, so ratios in quadrant III and quadrant I are positive but have very different meanings. Not a big problem except that you need to be careful about it
- 2) Ratios do not provide a clear idea of the size of the programs; not a big problem either. Hopefully the study provides a good description of the program
- 3) It is not trivial to calculate the confidence interval of ratios (this is a problem **if** you have individual-level data)
- Because of the CI problem, Stinnett and Mullahy (1998) proposed an alternative to ICER: **Net Health Benefits (NHB)**
- But doing so clarified the meaning of the ICER

First, Net Monetary Benefit

- The idea of NMB is based on basic algebra:
- The decision rule is $\Delta C/\Delta E < R_T$; same as $ICER < R_T$
- Do some algebra and you get $R_T \Delta E - \Delta C > 0$
- $R_T \Delta E - \Delta C$ is called the **net monetary benefit**
- If net momentary benefit is positive, the intervention is cost-effective
- Nothing magical about this, just a re-arrangement. **But it does change the measurement units (!)**
- $(R_T \times \Delta E)$ is \$ and so is ΔC
- So net monetary benefit is expressed in dollars. Essentially, with a little bit of algebra we have translated CEA into cost-benefit. Note the key step: we are using a threshold to make a decision and we make the decision about \$\$

Value of Life (**Wake up!**)

- The threshold R_T is measured in the same units as the ICER
- For example, a common threshold is about $\$150,000$ per QALY. So it's $\frac{\$}{QALY}$
- If the measure of effectiveness are in QALYs, one interpretation of the threshold is that it's the "value" of 1 QALY: it's the **value of one year of life** (adjusted for quality)
- If the ICER of an intervention is, say, $\$500,000$ per QALY, we would say that it's not cost effective. The max we are willing to pay for 1 QALY is $\$150,000$
- Got it? The threshold is the value of life!
- In a future class we will talk about the **statistical value of life**. Essentially, ways in which we could come up with that number (**stated preferences, revealed preferences**). **So it would be a way to come up with a threshold**

Second, Net Health Benefits

- We can keep doing algebra to rearrange the decision rule $\Delta C/\Delta E < R_T$ again:
- $\Delta E - \frac{\Delta C}{R_T} > 0$, which is called the **Net Health Benefit**
- So now the decision rule is that the incremental gain in effect has to be greater than the incremental cost over the threshold value
- This changes the units again. **Now the comparison in terms of effect units, not money**
- For NHB to be positive, the health gain has to be greater than that of investing the same resources in an alternative with the cost-effectiveness R_T
- So now you can view the threshold R_T as the **opportunity cost**. If an ICER is greater than the threshold, we are saying that our money would be more effectively used in a different, more cost-effective alternative

So why are we doing algebra?

- I did the algebra because:
  1. This is a good way to introduce the big ideas about the use of a threshold value
  2. Your textbook explanation is not very good
- Stinnett and Mullahy (1998) did it because the rearrangement transforms the decision rule based on ICER into a linear function for which building confidence intervals is easier
- If the algebra doesn't help you, the conclusion is: using a threshold value as a decision rule is **equivalent to translating benefits into money** and the **threshold is also an opportunity cost**
- **Remember the part of decision rule transformed into money when we cover cost-benefit analysis**

Thresholds, where do they come from?

- Where does the $\$100,000$ to $\$150,000$ number comes from? In short, the answer appears to be that the number **came from nowhere**. For many years, it was considered to be $50K$
- 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**
- Really, **since when the US government uses cost-effectiveness to make decisions??**

So where did it come from?

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

Table 1. Steps in the development of the $50,000 per quality-adjusted life-year cost-effectiveness threshold.

Year	Action	Ref.
1992	First widespread use of CE thresholds, $20,000 and $100,000 per QALY, in Canada	[33]
1992	First publication to use a CE threshold of $50,000 per QALY	[87]
1995	Second publication to use CE threshold of $50,000 per QALY and first to promote it	[92]
1995	Widely cited article using CE threshold of $50,000 per LY published	[61]
1996	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	[127,128]
1996	Use of $50,000 per QALY or LY as a common point of comparison begins to take off	[70,76,95]
1997–1998	Experts begin referring to anything costing less than $50,000 per QALY as 'economically attractive' [75] or 'reasonably efficient'	[99]

CE: Cost-effectiveness; LY: Life-year; QALY: Quality-adjusted life-year.

In short...

- 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**: $50,000 * (1.03)^{25} = \$104,689$
- Medical inflation is a lot higher: $50,000 * (1.045)^{25} = \$150,272$
- *Notice the first funny thing*: the 50K **comes from nowhere but it is close to what we use now adjusted for inflation**
- 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)

Other thresholds

- From Neumann et al (2014)

Cost-Effectiveness Thresholds Referenced by Authors of U.S.-Based Cost-Utility Analyses, 1990–2012.*

Threshold	1990–1999 Analyses (N = 207)	2000–2009 Analyses (N = 851)	2010–2012 Analyses (N = 444)
		percent	
$50,000 per QALY	19.3	36.6	36.9
$100,000 per QALY	6.3	7.8	16.9
Both $50,000 and $100,000 per QALY	3.9	19.9	23.7
Other	18.4	10.6	7.4
No threshold referenced	51.9	25.1	15.3

* 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. So that could be one way
- 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**?
- 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
- *Notice the second funny thing*: It's very close to the the $50K adjusted for inflation

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)
- *Not a third funny thing* but that threshold is likely to be too high: there is not much connection in the US between spending more and health
- 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**

Other standards

- Different organizations have different standards. For example, for drugs (Neumann and Cohen, 2015)

Institute for Clinical and Economic Review (ICER)	Incremental cost-effectiveness plus care value components Comparative clinical effectiveness Other benefits and disadvantages Contextual considerations Budget impact	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

American College of Cardiology–American Heart Association (ACC–AHA)	Clinical benefit vs. risks Magnitude of net benefit Precision of estimate based on quality of evidence Value (cost-effectiveness)	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.

Summary

- 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