Outcome Measures

David Torgerson
Director, York Trials Unit

Background

- There are numerous methods of measuring outcomes in trials.
- Usually, need to measure 'clinical' effects and quality of life.
- Often quality of life and clinical measures will correlate but may not.

'Clinical' Outcomes

- These are numerous and are often 'surrogates' for 'real' outcomes.

Surrogate vs Real measures

	Surrogates	Real
Vascular Disease	Blood pressure, lipids	Stroke, angina, heart attack, death.
Osteoporosis	Bone mass, bone turnover	Fracture
Partner assault	Changes in q'naire	Reduction in assaults.
MSc lectures	Enjoyment, satistfaction	Knowledge?

Problems with surrogates

- Change in surrogates may not lead to changes in real outcomes.
- Sodium flouride INCREASES bone mass but also INCREASES fractures.
- HRT reduces cholesterol levels but INCREASES risk of stroke and cardiac disease.

HRT

- HRT profoundly affects a wide range of surrogates. Improves blood cholestrol increases blood flow to brain.
- Trials with REAL outcomes shows increases in deaths due to cardiovascular disease and increased incidence of dementia.
- Does increase bone mass and reduce fractures (only 1 surrogate was correct).

AIDS

- Some successful anti-AIDS drugs have little or no effect on cellular markers of disease progression. BUT in trials of the drugs with AIDS death as the outcome they did reduce deaths.

Satisfaction

- Some trials show either qualitatively or quantitatively an improvement in treatment satisfaction but no change in real outcome.
- Example, counselling for women after traumatic childbirth increases satisfaction with the service but also INCREASES post natal depression.

CBT on employment

- A RCT of the use of CBT on the rate of finding a job showed no difference between the groups in 'job seeking activities' (e.g., number of interviews, number of job applications etc) BUT the trial showed those allocated to CBT were significantly MORE likely to get work (34%) than the controls (13%) ($p < 0.001$).

Proudfoot et al. Lancet 1997;350:96-100.

CBT & employment

- Had the trial only measured job seeking behaviour then we would have concluded, erroneously, that CBT was a useless intervention at increasing employment for the long term unemployed.

Atkins Diet

- Dieticians dislike the Atkins Diet at it goes against 'accepted' wisdom. HOWEVER, whilst weight loss isn't much different from a low carbohydrate diet lipids (surrogates) for cardiovascular disease are better.
- It seems surrogates are mistrusted if they go against accepted wisdom but trusted if they confirm the prior hypothesis.

Why use surrogates?

- If surrogate markers are misleading why use them?
- Often cost – real outcomes of death or disablement require huge expensive trials markers will tend to confirm that a drug is acting as theory suggests it should. Example, bone mass changes confirm drug is reaching the bone and exerting an effect.

Class effects

- Often 'me to' drugs use markers as they act in a very similar way as established treatments and the assumption is made that they if they reduce the surrogate they will also reduce the real event.
- Example, daily bisphosphonate treatment increases bone mass and reduces fractures. Weekly treatment increases bone mass the ASSUMPTION is that weekly will reduce fractures as much as daily.

Sample size

- Usually surrogates need a much smaller sample size to show an effect, which reduces the cost, increases the speed of the trial etc.
- However, need to be wary of their use.

Quality of Life

- The aim of most health care is to improve quality of life.
- For many people extending life or preventing death is not necessarily the most important aspect.

Quality of Life

- Many treatments will extend life or increase the probability of survival but at the 'expense' of very poor quality of life. For example, radical surgery of head and neck cancer will improve survival from very low levels by only a small amount. Terrible quality of life effects: patient can't speak properly; difficulty eating, terrible disfigurement. The majority of patients will still die but have their remaining life span of very poor quality.

Measuring quality of life

- A number of quality of life scales are widely used:
  - Disease specific;
  - Generic measures;
  - Utility measures.

Disease specific

- These are questionnaires that will ask specific questions relating to the health condition. For example, the Roland & Morris back pain scale asks 24 questions about disability related to your back (e.g., do you have trouble getting out of a chair because of your back pain?)

Disease specific measures

- These measures have a number of advantages in that they are 'sensitive' to changes in the condition. BUT they will not pick up other general health disadvantages or benefits of treatment.
- For example, will not pick up cessation of depression through curing back pain.

Generic measures of health

- These have questions asking about general health (e.g., SF36; SF12; Nottingham health profile (NHP) Women's Health Questionnaire).
- Advantages in that they will pick up other effects of treatments.
- Disadvantage may not be sensitive to small, but important, health effects.

Example of SF36

Your Feelings

(Please Circle **One Number on Each Line**)

- These questions are about how you feel and how things have been with you **during the past month**. For each question please give the one answer that comes closest to the way you have been feeling. How much during **the last month**:	All of the time	Most of the time	Some of the time	A little of the time	None of the time
a) Have you felt calm and peaceful?	1	2	3	4	5
b) Did you have a lot of energy?	1	2	3	4	5
c) Have you felt so down in the dumps that nothing could cheer you up?	1	2	3	4	5

SF36/SF12

- The Short Form (SF) 36 item questionnaire is one of the most widely used generic quality of life instruments. It is derived from the much longer Medical Outcome Survey Instrument developed by the RAND corporation as part of a massive RCT of payment systems for health care treatment.

SF36 domains

- The SF36 has 8 domains;
  - Physical functioning
  - Social functioning
  - Role physical
  - Role emotional
  - Mental health
  - Vitality
  - Bodily pain
  - General health

Two or eight?

- The eight domains can be collapsed into two domains
- Physical and Mental health.
- Advantages of just two domains include:
  - Less likely to have a Type I error;
  - More resistant to missing items.

SF12

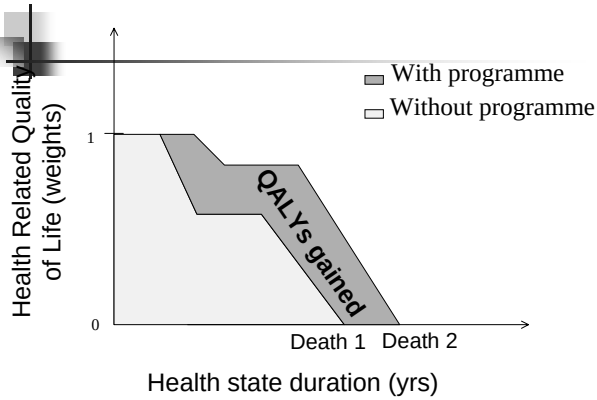
- Only 12 questions, scored into the two domains of mental and physical health.
- Advantages of a shorter questionnaire, less data entry, higher response rates.

Utility measures

- Problem will all of the other measures the scales do not have ratio properties. A person who scores 60 on the SF12 is better than someone who scores 30 but not twice as good. This makes it difficult to compare across conditions or use for economic analysis.
- Need a utility measure.

Quality Adjusted Life Years (QALYs)

Expressing impact using QALYs



Utility

- Another issue underpinning the valuation of utility measures is the concept of resource scarcity. Economists assign a value to something if one is willing to pay for it.
- Life has a value because people are willing to trade an increase risk in death to improve their utility (e.g., North sea divers have an enhanced salary to compensate them for increased risk of death).

Utility definition

- Numbers that represent the strength of an individual's preference for a particular health state under uncertainty.

Measuring utility

- One way of measuring and valuing utility is through a 'standard gamble'. Typically people are presented with a range of scenarios with assigned probabilities. Would you have surgery for your hip arthritis with a probability of dying of 0.01 or not? The values are varied until people accept surgery and this gives a mortality weight to the quality of life. This can be converted into monetary forms using wage differentials etc for risky occupations.

Standard Gamble - Problems

- No one understands it.
- Can produce 'illogical' answers (e.g., people choosing certain death for treatment for a minor illness).

Time Trade Off – a solution?

- An alternative widely used is a TTO.
- In this approach people are given a scenario such as "Imagine you have 10 years left to live in your current health state, how many of these years would you give up to be in perfect health?" The more that is given up the greater the quality of life gained by treatment.

TTO - problems

- Difficult to understand, many incorrect answers.

Willingness to pay

- In this approach people are asked about their willingness to pay for a treatment given an outcome scenario. For example, women were asked their WTP for HRT for the treatment of severe menopausal symptoms. Most were WTP significant amounts (substantially exceeding the cost) for treatment.

WTP - problems

- Usually the answer is £25 whatever you ask or people refuse to give an answer or say an infinite amount.
- On a practical side if you put WTP questions in your questionnaire you get letters sent to MPs with accusations that it is plot to 'privatise' the much loved NHS.

Non-response to WTP

- A study compared the response rates to two methods of eliciting preference WTP and Willingness to Wait. 15% of participants answered WTW but NOT WTP, whilst only 3% would answer WTP and not WTW (79% answered both).

Thomas et al, JHSRP, 2000;5:7-11.

Conjoint Analysis

- This is an approach originally used in Environmental economics. Patients are given scenarios of a health care intervention and asked to choose (e.g., you get the operation in a month but have to go to a hospital 100 miles away or get it in 8 months at your local hospital).
- Patients are then asked to choose between scenarios and a utility can be derived between health care scenarios.

CA - Problems

- Difficult questionnaire can lead to misunderstandings.
- Currently an approach that is generating a lot of PhDs.

What is the ideal?

- A simple questionnaire that is sensitive and reliable and produces are utility weight for quality of life.
- Not there yet – BUT there are some questionnaires that try.

Utility measurements

- Several available (e.g, EuroQol, HUI) what they all CLAIM to is to produce a ratio scale.
- Their main disadvantage is they are very insensitive to changes in health status.

The EQ-5D

Mobility

I have no problems in walking about ❑
I have some problems in walking about ❑
I am confined to bed ❑

Self-Care

I have no problems with self-care ❑
I have some problems washing or dressing myself ❑
I am unable to wash or dress myself ❑

Usual Activities (e.g. work, study, housework, family or leisure activities)

I have no problems with performing my usual activities ❑
I have some problems with performing my usual activities ❑
I am unable to perform my usual activities ❑

Pain/Discomfort

I have no pain or discomfort ❑
I have moderate pain or discomfort ❑
I have extreme pain or discomfort ❑

Anxiety/Depression

I am not anxious or depressed ❑
I am moderately anxious or depressed ❑
I am extremely anxious or depressed ❑

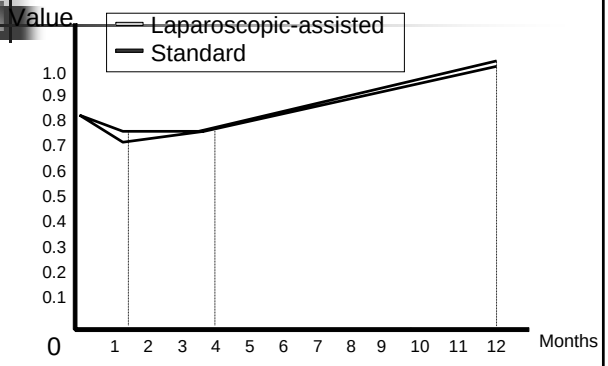
EuroQol Scoring

- Scores from 0 (dead) to 1 perfect health. Also allows negative scores (health states worse than death, e.g., a short holiday in Disneyland Paris).

Valuing the EuroQol

- There are 245 'health states' in the EuroQol (including negative ones). A large survey using TTO has attached a utility weight to each.

Health Outcomes - QALYs



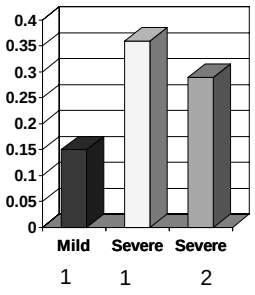
EuroQol problems

- The EuroQol is VERY insensitive to quite large changes in quality of life.
- Large gaps in the scale (e.g., cannot score between 0.88 and 0.99).
- Statistically has undesirable properties. Not normally distributed heavily skewed towards higher values.

Whose preferences?

- There is debate on how to measure a given health state. Should patients in that health state value it? Or should people not in the health state give it a value.
- Who to choose matters as generally people in a health state do not value it as badly as people not in a health state.

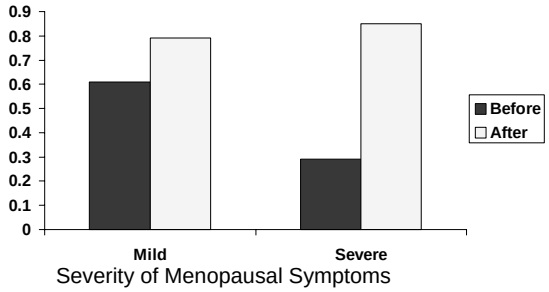
Disutility of menopausal symptoms



1 Daly et al, BMJ 1993;307:836-40

2 Zethraeus. Health Economics 1998;7:31-8.

Utility gain of HRT



Daly et al, BMJ 1993;307:836-40

What to use?

- Generally, should use a condition specific measure; general measure and utility measure as well as 'clinical' measure of outcomes.

Back pain Trial

- In back pain trials we used the following:
  - EuroQol (for economic evaluation);
  - Roland & Morris Back pain scale;
  - Aberdeen back pain scale;
  - SF36

York Backpain Trial

- In the York back pain trial we found significant differences in favour of the intervention in the Roland & Morris and Aberdeen back pain scale but non-significant differences in the EuroQol.
- Reason? EuroQol relatively insenstive to changes in small but important measures of outcome.

York Back Pain Trial

	Control	Intervention	Diff	p
R&M	-1.77	-3.19	1.42	0.02
Aberdeen	-8.48	-12.92	4.44	0.01
EQ5D	0.089	0.111	-0.02	0.47

Klaber-Moffett et al, BMJ 1999;319:279.

Cost and QoL

- Economic evaluations want to calculate a cost utility ratio. A EuroQol gain of 0.02 for a cost of £600 or lower is likely to be cost effective (I.e., <£30,000 per QALY), BUT difference is nowhere near statistically significant!
- Is this a Type II error? Possibly as both the 'condition specific measures' showed a significant improvement.

Cost and QoL

- A disadvantage of QALY is its inability to value short intense pain. Consider a local anaesthetic for removing your toenail. Let's assume it costs £5 for the anaesthetic and labour costs. The alternative is for it to be removed without. Assume QoL is 0 for 1 hour after removal. The QALY gain from an LA = 0.000114155, divide this into the cost and the ratio = £44,000 – NOT cost effective!
- This is nonsense as the WTP for virtually anyone considerably exceeds £5 for an LA.

Battery of measures

- Generally we use a battery of outcome measures:
  - To measure all relevant domains;
  - Also helpful if one is using an 'iffy' outcome measure.

Venus I trial

- In this trial we used the following measures:
  - SF12
  - Euroqol
  - Hyland condition specific measure for leg ulcers.

What did we find?

- We found that the Hyland measure was very insensitive. It did correlate with ulcer severity but less so than the SF12.
- There was a lot of missing data from the scale.
- We, therefore, used the SF12 as the main QoL measure.
- In other Venus trials (II & III) we are not using the Hyland.

Hyland

- However, because we had used a relatively sensitive measure of general outcome (SF12) we could still look at Quality of Life.

COLPO outcomes – overkill?

- In the MRC COLPO trial for urinary incontinence the following are used:
  - Pad test (jump up and down after drinking a litre of water);
  - Bristol Symptom Q'naire;
  - King's symptom q'naire;
  - Urinary distress inventory;
  - SF36;
  - EuroQol;
  - Sabbattsberg sexual rating scale.

Why so many?

- Partly urinary incontinence affects many aspects of health but additional urinary incontinence q'naires included because referees couldn't agree which one we should use and the PI decided, for a quiet life, to include them all.

Identifying QoL measures

- There are huge numbers of QoL measures for nearly every conceivable condition. Website at Oxford has a database of QoL instruments. Often clinicians will develop their own – not generally recommended as it is a specialised task, needing psychologists and statisticians.

What makes a good QoL measure?

- Appropriateness to the research question.
- Reliability (low random error)
  - Internal consistency
  - Reproducibility
- Validity (face and construct)
- Responsiveness.

QoL measures

- Precision (sensitive to changes)
- Interpretability
- Acceptability
- Feasibility.

Some statistical properties

- Need a measure to avoid 'ceiling' and 'floor' effects. Some measures have a floor effect cannot measure really poor quality of life and vice versa.
- A population at baseline that either scores nearly the maximum or minimum on a measure the wrong measure is being used.

Conclusions

- Need to identify outcomes that are of interest to the patient NOT the clinician, biologist or social scientist.
- Surrogate outcomes can mislead.