

Interpretation of Health-Related Quality of Life - HRQOL -

Madeleine King

President ISOQOL 2007
Biostatistics & Outcomes Research
Centre for Health Economics Research & Evaluation
University of Technology, Sydney, Australia

Workshop
2nd Turkish National HRQOL Congress
5 April 2007 Izmir Turkey

This presentation is based on a book chapter by
David Osoba and Madeleine King

Interpreting QOL in individuals and
groups: meaningful differences

In: *Assessing Quality of Life in Clinical Trials*

Peter Fayers & Ron Hays (Eds)

Oxford University Press

2005

Overview

- Background to HRQOL assessment
 - definition, motivation, scoring & analysis, missing data
- Types of “differences” in HRQOL that need to be interpreted
- Approaches to interpretation

“HRQOL” - Definition

- “QOL” is very broad, includes more than effects of health on well-being
- Health care context: the effect of disease and treatment on well-being
 - Health-related
 - Subjective, self-assessed
 - Multi-dimensional – physical, social, mental, and others
- The individual’s perception of the effect of disease and treatment on his/her ability to function physically, socially and emotionally
- Still a broad umbrella
 - health status, functional status, subjective well-being, patient-reported outcomes (PROs)

QOL v PRO?

- FDA (USA) Guidance for Industry
 - **Patient-Reported Outcome** Measures: Use in Medical Product Development to Support Labeling Claims
- PRO
 - any aspect of a patient's health status that comes directly from the patient
- Used in clinical trials
 - To measure the impact of an intervention
 - on one or more aspects of patients' health status
- PRO concepts
 - purely symptomatic, eg pain
 - more complex concepts, eg Activities of daily living (ADL)
 - extremely complex concepts, eg QOL

Motivation

Why assess HRQOL?

- Empirical evidence about the effects of treatment on HRQOL can help patients, health care professionals and policy-makers to make better informed choices among health care options

Fundamental questions

- Why is HRQOL measurement relevant?
- What aspects of HRQOL are of interest or are likely to be affected by disease or treatment?
- Which instrument(s) should be used?
- How will the results be analysed? presented? interpreted? used to inform practice or policy?
- Answers are specific to each situation

Interpretability

- An attribute of an instrument, along with validity & reliability LOHR et al 1996
 - “the degree to which one can assign qualitative meaning - that is, clinically or commonly understood connotations - to quantitative scores”
- Like validity, interpretability is not established by a single study - it develops gradually as a body of evidence accumulates with repeated experience from a variety of perspectives

Interpretability

- We need to understand and explain HRQOL differences in ways that are relevant to the health care decision at hand and the people who are involved in that decision
- There are a number of approaches, but before we can interpret scale scores we need to understand the process of measurement

What is measurement?

“assigning numbers by rules”

Stevens (Science 1946)

Questionnaires have simple rules for assigning numbers

- Standard set of questions
 - a representative sample of all relevant items
 - operationalises the definition of HRQOL domain
- Standard set of response scales
 - assign numbers to perceptions
- Standard scoring algorithm
 - defines the measurement scale
 - each dimension score is the (weighted) sum of item scores
 - sometimes linear transform: eg 0-100 scale range
 - sometimes norm-based: (mean, SD) eg (50, 10)

E.g., SF-36 Physical functioning 10 questions

- Vigorous activities – running
- Moderate activities – vacuuming
- Lifting or carrying groceries
- Climbing several flights of stairs
- Climbing one flight of stairs
- Bending, kneeling, stooping
- Walking more than a kilometer
- Walking several blocks
- Walking one block
- Bathing or dressing

Work out your SF-36 physical functioning score

10 items

1. Vigorous activities – running
2. Moderate activities – vacuuming
3. Lifting or carrying groceries
4. Climbing several flights of stairs
5. Climbing one flight of stairs
6. Bending, kneeling, stooping
7. Walking more than a kilometer
8. Walking several blocks
9. Walking one block
10. Bathing or dressing

Response options:

- | | |
|------------------------|---|
| Yes, limited a lot | 1 |
| Yes, limited a little | 2 |
| No, not limited at all | 3 |

Formula & example for transformation of raw scale scores

SF-36 Manual & Interpretation Guide page 6:18, <http://www.sf-36.org>

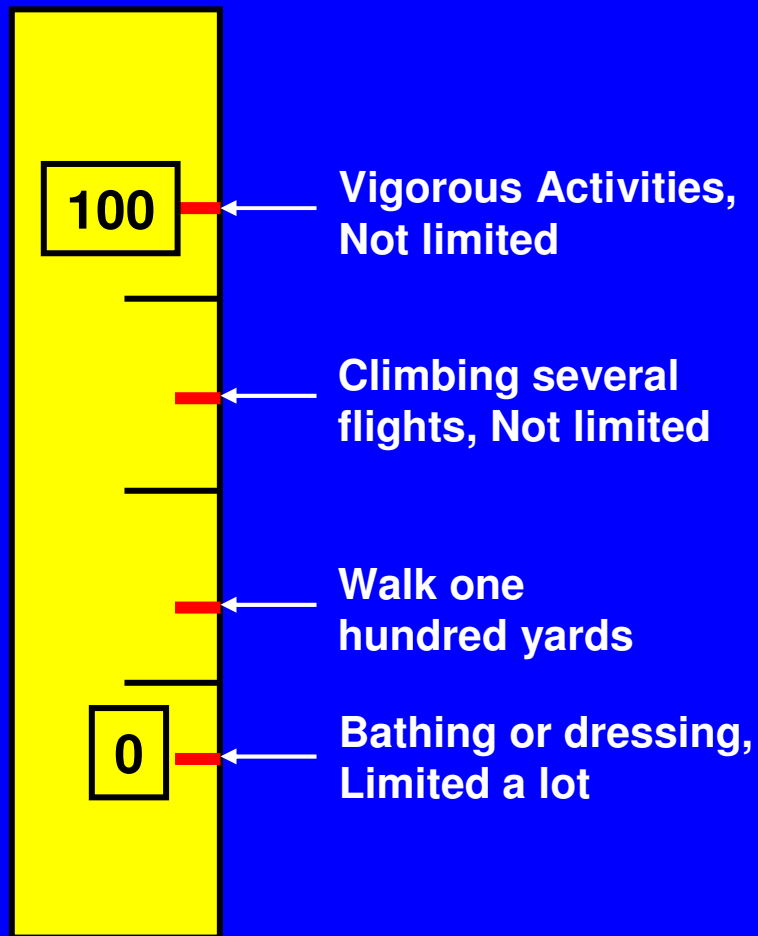
Transformed scale =

$$\frac{\left[\text{actual raw score} - \text{lowest possible raw score} \right]}{\left[\text{possible raw score range} \right]} \times 100$$

Example: A physical functioning raw score of 21 would be transformed as follows:

$$\frac{\left[21 - 10 \right]}{\left[20 \right]} \times 100 = 55$$

Physical Functioning “Ruler”



Source: Haley et al, JCE 1994

Scoring procedures

- Turning responses to individual questions into scale scores, i.e., variables for statistical analysis
 - single items or multi-item scales?
- Does the instrument have standard scoring procedures?
 - If so, use them - facilitates interpretation via comparability across studies that use the instrument

HRQOL is a “subjective measure”

- used by some as a pejorative term
 - soft, rubbery, not valid or reliable. WRONG!
- HRQOL is:
 - Subjective phenomenon
 - Objective measurement
 - psychometric heritage - eg, intelligence, depression
 - methods to assess validity & reliability of scales
- An important distinction!

Why is interpretation hard?

- measuring a “latent” (unobservable) thing
- respondents’ perceptions, subjective
- “inconsistency” in how response scales are used
 - among patients & within patients over time
- scales are ordinal
 - values on scales are arbitrary
- lots of different scales – every scale is different
 - little familiarity
- few instruments have interpretation guidelines
 - SF-36 is notable exception
- significance may differ with perspective
 - Eg, patients, clinicians, policy-makers

But despite all of that ...

- HRQOL “numbers” behave quite sensibly:
 - People & groups with better health tend to have higher mean scores
 - (good!) instruments register increases in mean scores of people & groups whose health really improves
- Familiarity with a scale’s behaviour is the key to interpreting outcomes measured on it
- There are a number of approaches ...
 - they are all useful in different ways & contexts
 - Lets think first about the types of differences we try to interpret

4 types of differences in HRQOL

	Cross-sectional	Longitudinal
Individual	<p>Screening for people with high levels of symptoms / psych problems, who will benefit from targeted interventions.</p> <p>BEWARE: misclassification error</p>	<p>Monitor -benefits & side-effects of therapy – need to be confident that changes in scores observed on scales reflect true change.</p> <p>BEWARE: measurement error</p>
Group	<p>Describe & compare health of different populations – usually population health and clinical research applications. Instruments need to be SENSITIVE</p>	<p>Evaluating impact of interventions. More powerful than X-sectional comparisons. Instruments need to be RESPONSIVE</p>

Various Approaches to Interpreting HRQOL Results

Osoba & King (2005)

- Internally-referenced:
 - derived solely from the observed HRQOL scores in the primary data set and/or the instrument's known measurement properties
- Externally-referenced:
 - rely on information that is additional, or external, to the HRQOL data in the primary dataset
 - other well-established and understood variables measured in the primary dataset
 - HRQOL scores of the same instrument measured in other well-defined and understood samples

Internally-referenced approaches:

1. Content-based interpretations

Literal interpretations

Eg SF-36 Physical functioning

10 items

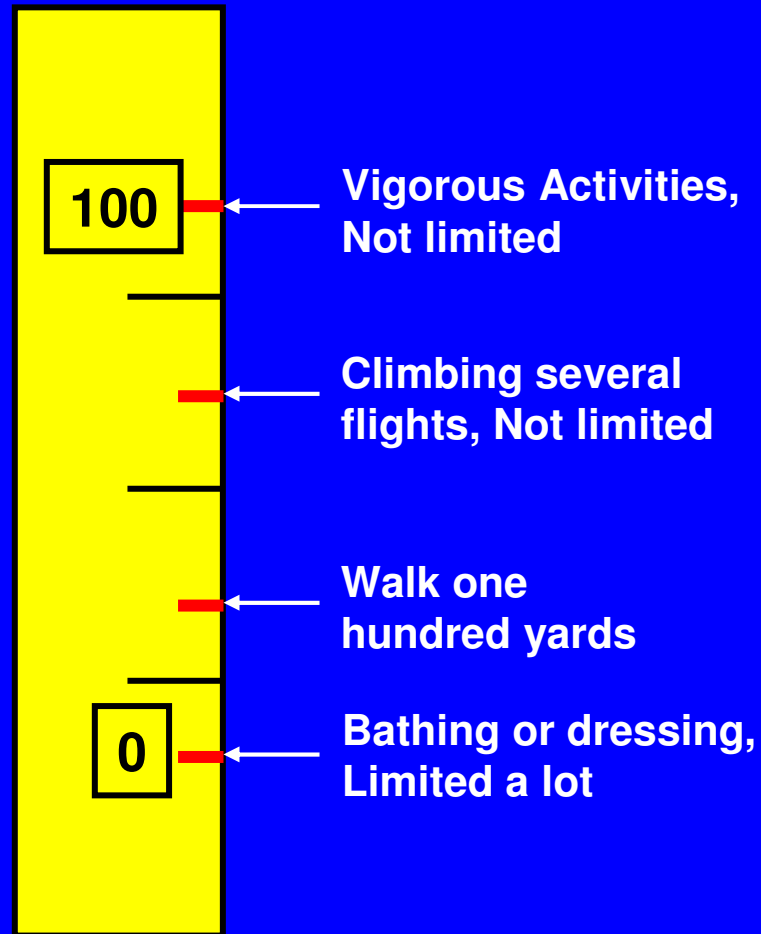
1. Vigorous activities – running
2. Moderate activities – vacuuming
3. Lifting or carrying groceries
4. Climbing several flights of stairs
5. Climbing one flight of stairs
6. Bending, kneeling, stooping
7. Walking more than a mile
8. Walking several blocks
9. Walking one block
10. Bathing or dressing

Response options:

- | | |
|------------------------|---|
| Yes, limited a lot | 1 |
| Yes, limited a little | 2 |
| No, not limited at all | 3 |

SF-36 Manual & Interpretation Guide Table 9.1

Physical Functioning ‘Ruler’



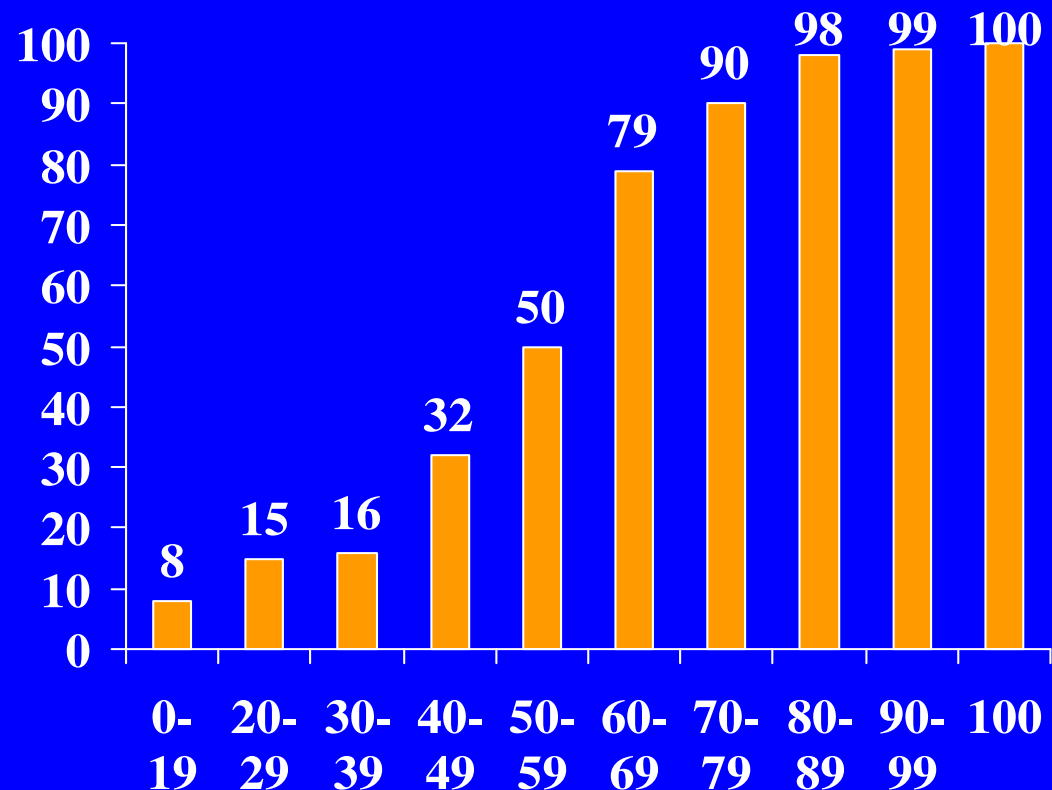
Source: Haley et al, JCE 1994

Literal Interpretations – strengths & weaknesses

- very easy and useful for single-item scales
- illustrate the abstraction that the numbers on the measurement scale represent
 - together with the item response scales, indicate the range of experience within the scale
- But specific rather than general interpretations:
 - eg “to walking 100 metres” vs “physical function”
- Only possible for end-points of the scale
 - all other scores may be reached by many combinations of items and responses
 - eg, there are 2850 possible ways to obtain a score of 70 on the SF-36 physical functioning scale

Another approach to content-based interpretation for multi-item scales

- Eg, As the SF-36 physical function scale scores increase from 40 to 50, 18% more people indicate they can walk one block without limitations



Internally-referenced approaches

2. Effect size

- $ES = \text{mean diff (HRQOL units)} / \text{SD (HRQOL units)}$
 - the size of the observed effect (MEAN DIFF) is interpreted in terms of variability among individuals (SD) - ie, “in SD units”
- since both “signal” (mean) and “noise” (SD) are measured on the same scale (QOL scale), their ratio is unitless
- standardised (unitless) measure allows comparison between different instruments (scales)

Cohen's Guidelines for Interpreting Effect Sizes

Statistical Power Analysis for the Behavioural Sciences 1988

- operational definitions
 - Small ES = 0.2 one fifth of a standard deviation
 - Medium ES = 0.5 half of a standard deviation
 - Large ES = 0.8 four fifths of a standard deviation
- “arbitrary conventions ... recommended for use only when no better basis for estimating effect size is possible” Cohen 1988; pages 12, 25.
- as collective experience accrues, it can be collated, synthesised & summarised
 - “evidence-based interpretation guidelines”

Evidence-based interpretation guidelines for effect-sizes to confirm or replace Cohen's

- For QLQ-30 – KING QOLR 1996; 5(6); 555-67 and KING QOLR 2001; 10(3): 278.
 - Systematic synthesis of evidence from 55 papers
 - Using clinical criteria to classify observed differences and effect sizes as small, medium and large
 - yielded guidelines similar to Cohen's
- For FACT-G KING, STOCKLER, CELLA, OSOBA et al, QOLR 2003; 12(7): 771.
 - Systematic synthesis of evidence from 71 papers
 - small ES ~ 0.3, medium ES ~ 0.7, large ES ~ 1.8
- Osoba estimated ES associated with small medium & large “subjectively significant difference” (SSD)
 - similar to Cohen's
- evidence is stacking up - Cohen's guidelines are probably reasonable ... roughly in the right ball-park ...

Strengths & weaknesses of effect size approach to interpretation

- definitely useful for determining sample size in the planning phase of a HRQOL study
- Weaknesses of ES
 - Thinking in standard deviation units may not be intuitively accessible for everyone, eg clinicians, policy makers
 - cannot be used to interpret individual scores, except in terms of how many SDs a person is from a particular mean (eg, a norm or reference value)

3. Statistical significance

- This is how hypothesis-testing works:
 - Question:
 - New treatment B - is it better than current tmt A?
 - Conduct a study to test “null hypothesis”:
 - $H(0)$: HRQOL (tmt B) = HRQOL (tmt A)
 - If $p < 0.05$ - statistically significant
 - Then reject null hypothesis
 - Conclude HRQOL (tmt B) is better than HRQOL (tmt A)

But ...

- How much better does HRQOL have to be to convince patients, clinicians or policy makers to switch to Treatment B as the new best treatment?
- This difference is the “**clinically important difference**”
 - should be specified *a priori* to determine the appropriate sample size for the study to have the **power** to detect **the alternative hypothesis**

Clinical v statistical significance

- If, at the planning stage, a study is properly **powered** (ie the sample is large enough to detect the smallest **clinically important difference**) then the interpretation of its results is easy
- problems arise if sample size is based on other considerations, e.g., survival
- With a large enough sample, a tiny difference will be statistically significant, but have no practical or policy relevance

Bottom line

- Statistical significance can only be used to interpret group-based results if:
 - 1) the aim of the analysis is to test a hypothesis
 - 2) the sample size has been determined *a priori* on the *smallest clinically important difference*
 - Even then, the clinical significance should be considered and discussed to provide a useful interpretation of the results for readers
- Not useful for individual scores

Internally-referenced approaches

4. Standard Error of Measurement SEM

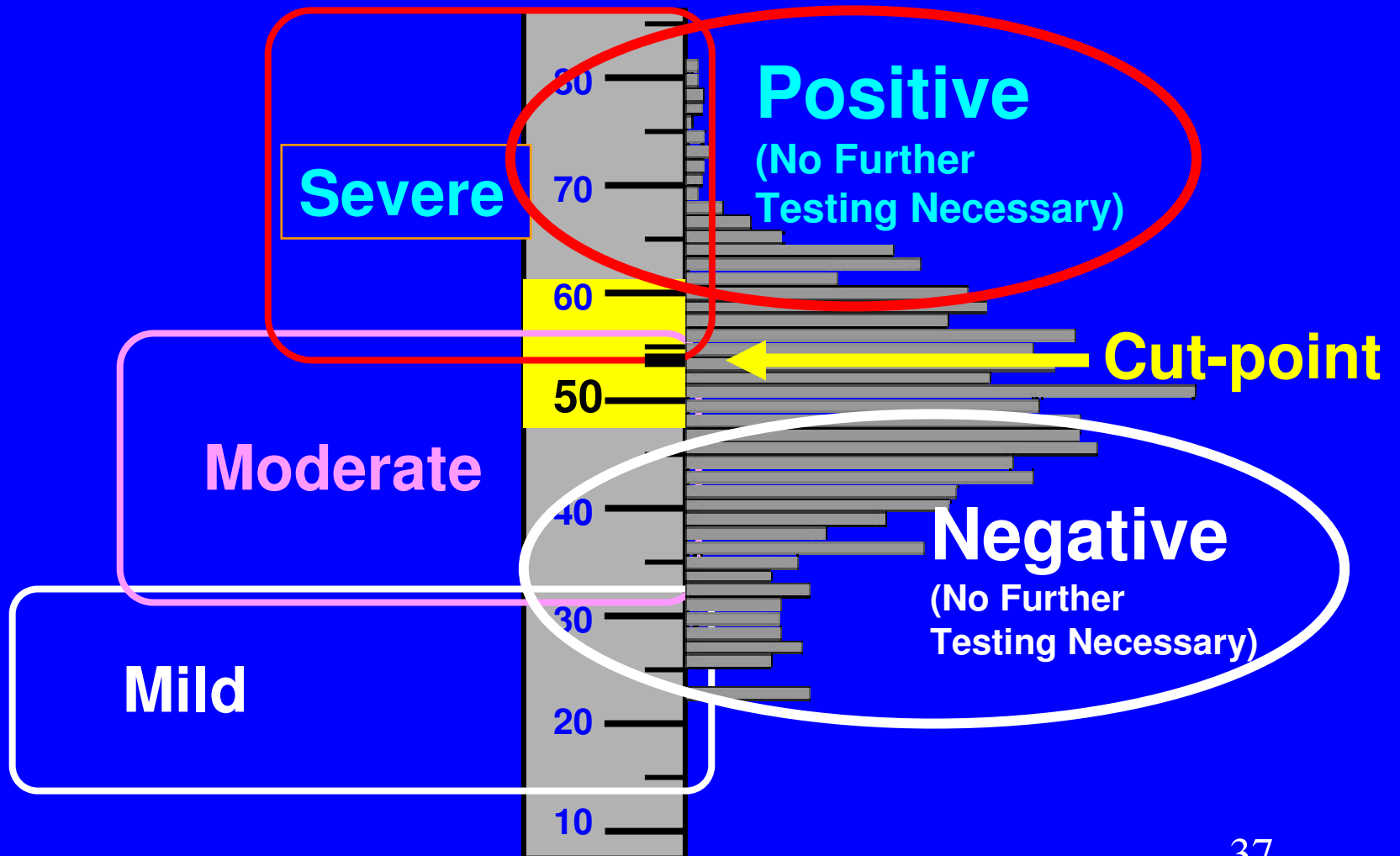
Anastasi & Urbina 1997, Wyrwich 1999 & 2002

- the most useful reliability estimate for individual-level applications
- indicates the smallest observed change that is likely to reliably reflect true change
- a function of scale reliability and between-person variance (ie, SD)

Interpreting an Individual Patient's HRQOL Scores

- use of HRQOL assessment to manage individual patients requires very precise measurement
- if using for screening, don't want to misclassify individuals
 - false negatives & false positives
- if using for monitoring over time, need to be sure “real change” has occurred

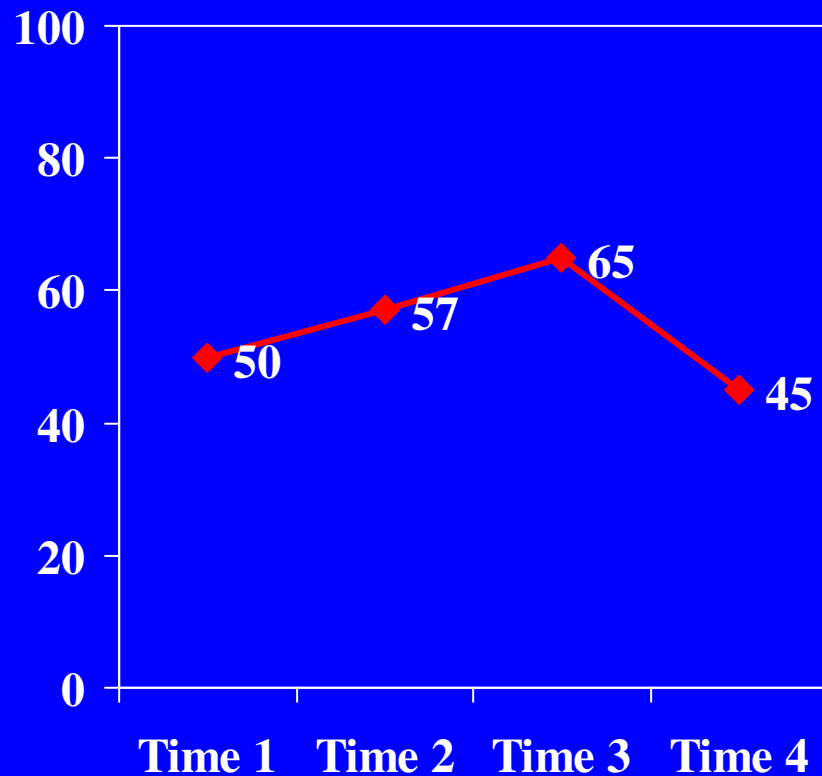
Why precision is needed for screening



Why precision is needed for monitoring change over time

- when we calculate a change in HRQOL between 2 times, it contains measurement error from both times
- so, unless we have very precision measures, how can we be sure that “apparent” change, ie, the change score on an imprecise instrument, reflects true change in HRQOL?

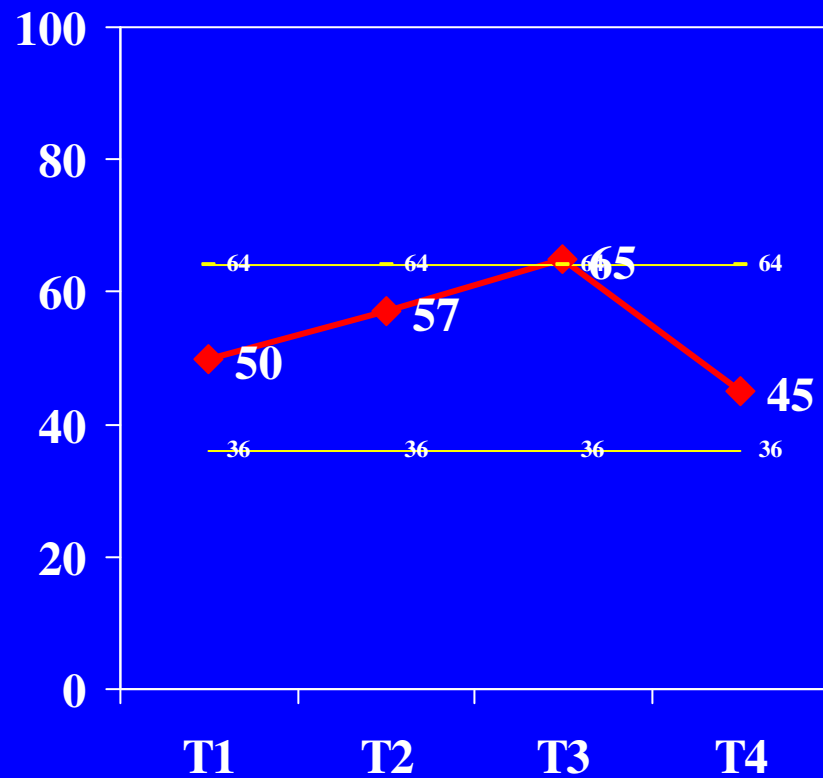
Consider this person's SF-36 physical function scores



- Is this evidence of improvement then deterioration?

SEM applied to the SF-36 physical functioning scale

McHorney and Tarlov QOLR 1995



- individual change score must be greater than 14 points (in either direction, improved or deteriorated) to reliably reflect true change in physical functioning

Surprising?

- ... this scale has 10 items and 21 possible scale values, and is certainly reliable enough for group-level research.
- similar conclusions for five commonly used health status instruments McHorney and Tarlov QOLR 1995
- we need more reliable instruments for the monitoring and management of individual patients

Review of internally-referenced approaches to interpretation

Derived solely from the observed HRQOL scores in the primary data set and/or the instrument's known measurement properties	Individuals		Groups	
	Cross-sectional	Longitudinal	Cross-sectional	Longitudinal
Content-based	✓	✓	✓	✓
Statistical significance			x	x
Effect size			✓	✓
Standard error of measurement	✓	✓		

✓ approach is useful

x approach is feasible but not appropriate

Blank cells indicate approach is not feasible, appropriate or useful

Externally-referenced Approaches to Interpretation

- rely on information additional, or external, to the HRQOL data in the primary data set
 - other well-established and understood variables measured in the primary dataset
 - clinical “known-groups”
 - HRQOL scores of the same instrument measured in other well-defined and understood samples
 - population and patient “known-groups”
 - norms and reference values

Externally-referenced approaches

1. Norms and Reference Values

- HRQOL scores from well-defined, representative samples provide “typical” scores
- Population norms are useful for generic instruments,
 - eg SF-36
- For condition-specific instruments, samples with particular stages of disease or on particular treatments provide useful reference values,
 - eg QLQ-C30 **Fayers et al EORTC 1998**

SF-36 Population Norms

USA

- SF-36 Manual & Interpretation Guide
Chapter 10 – USA Norms
- USA General population
 - males / females
 - age-groups
 - gender x age
- Health conditions
 - hypertension
 - congestive heart failure
 - diabetes II
 - acute myocardial infarction
 - clinical depression

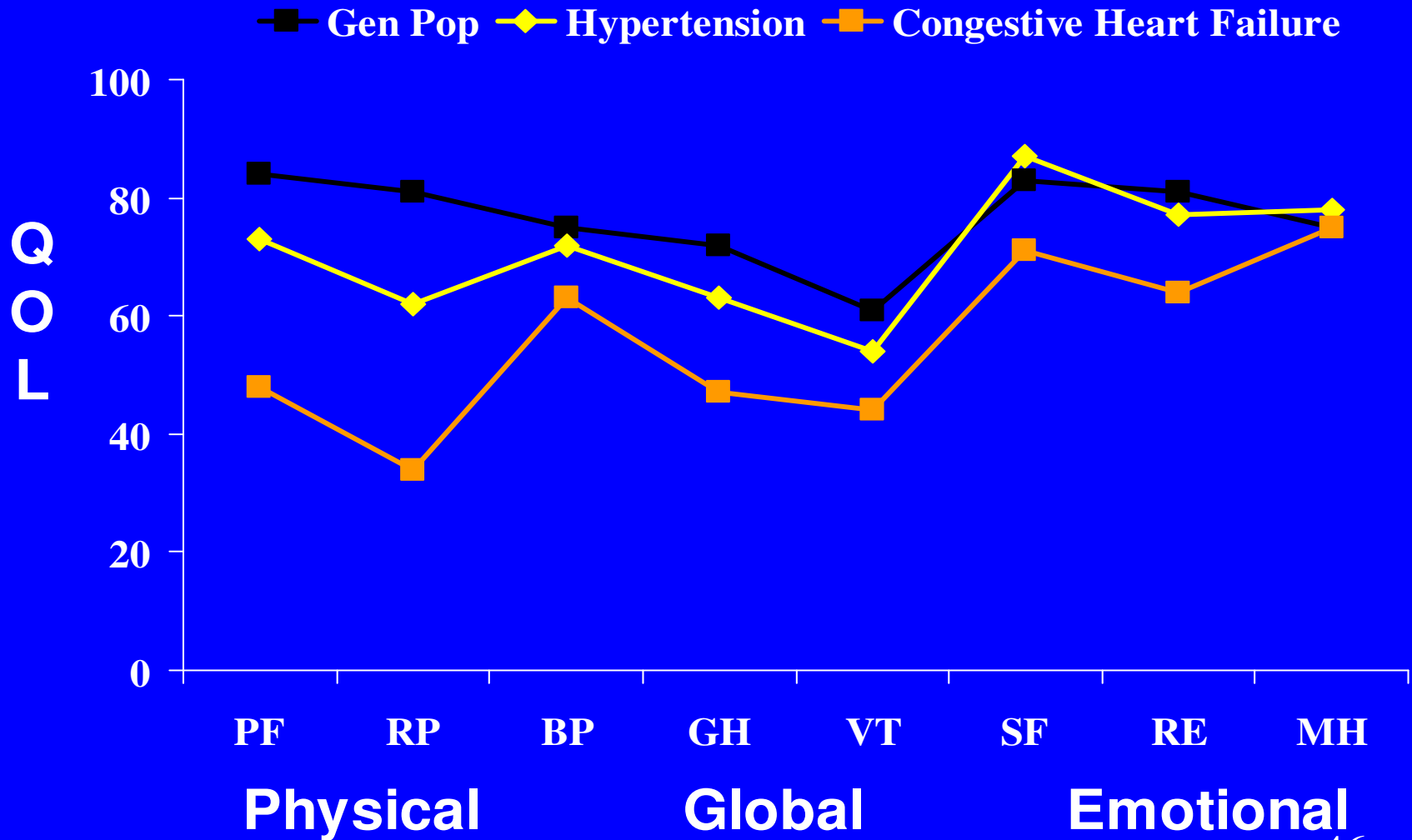
Other countries

Eg, Australia

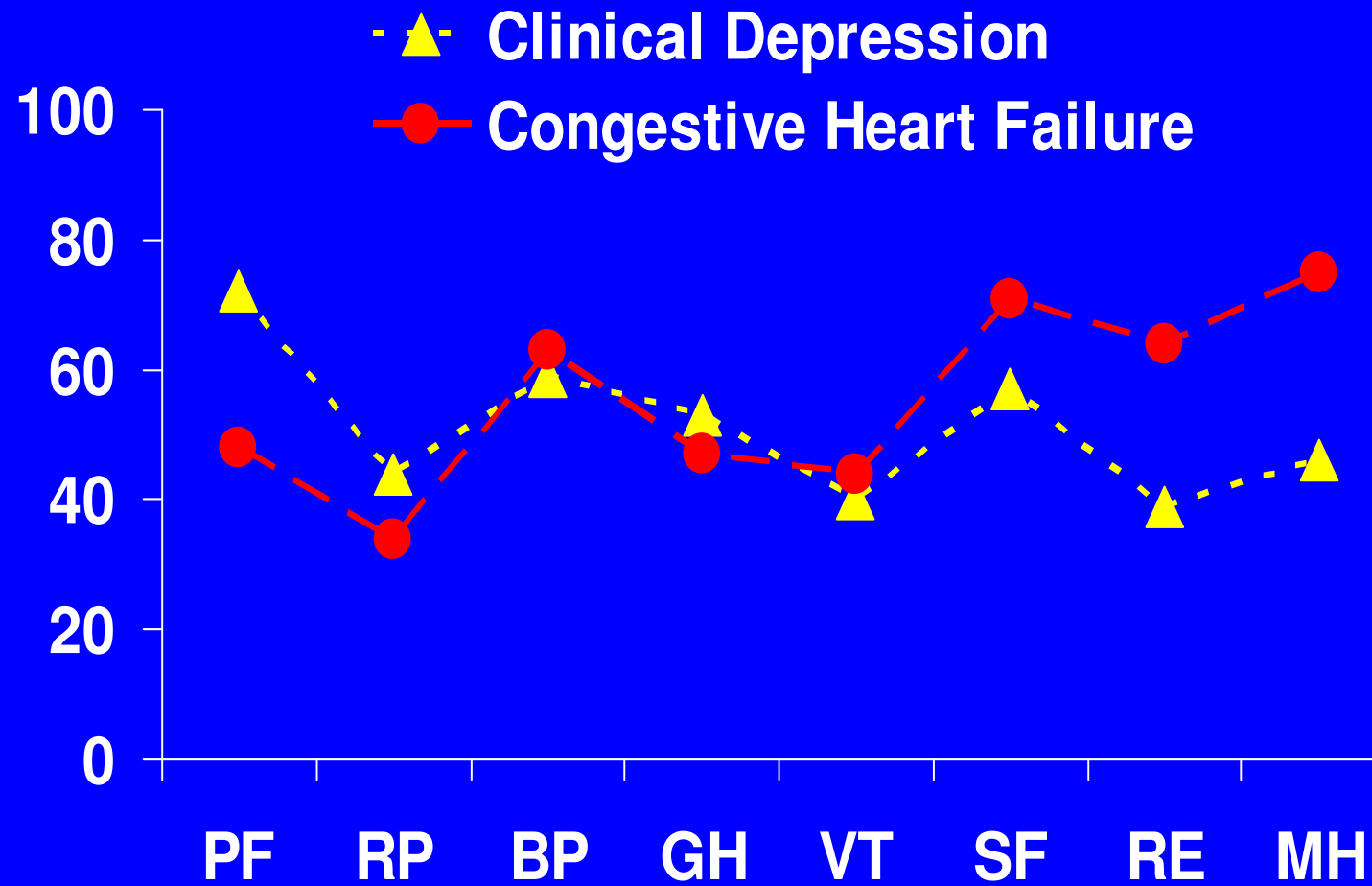
- ABS publication (currently out of print)
- 4399.0 National Health Survey:SF36 Population Norms, Australia
- ISBN 0 642 25720 5
- Australian General population
 - males / females
 - age-groups
 - gender x age

SF-36 "Profiles"

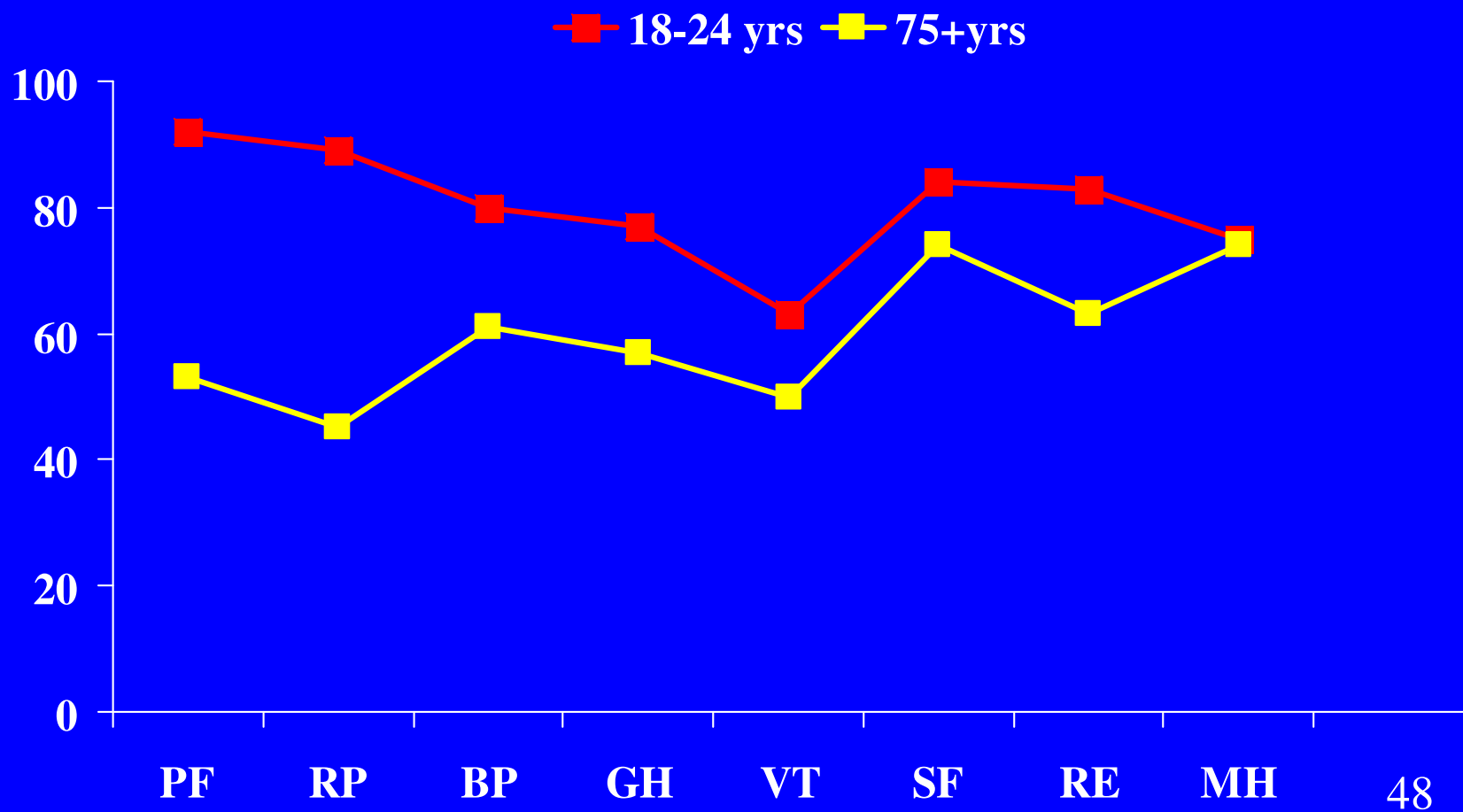
for general and medical populations



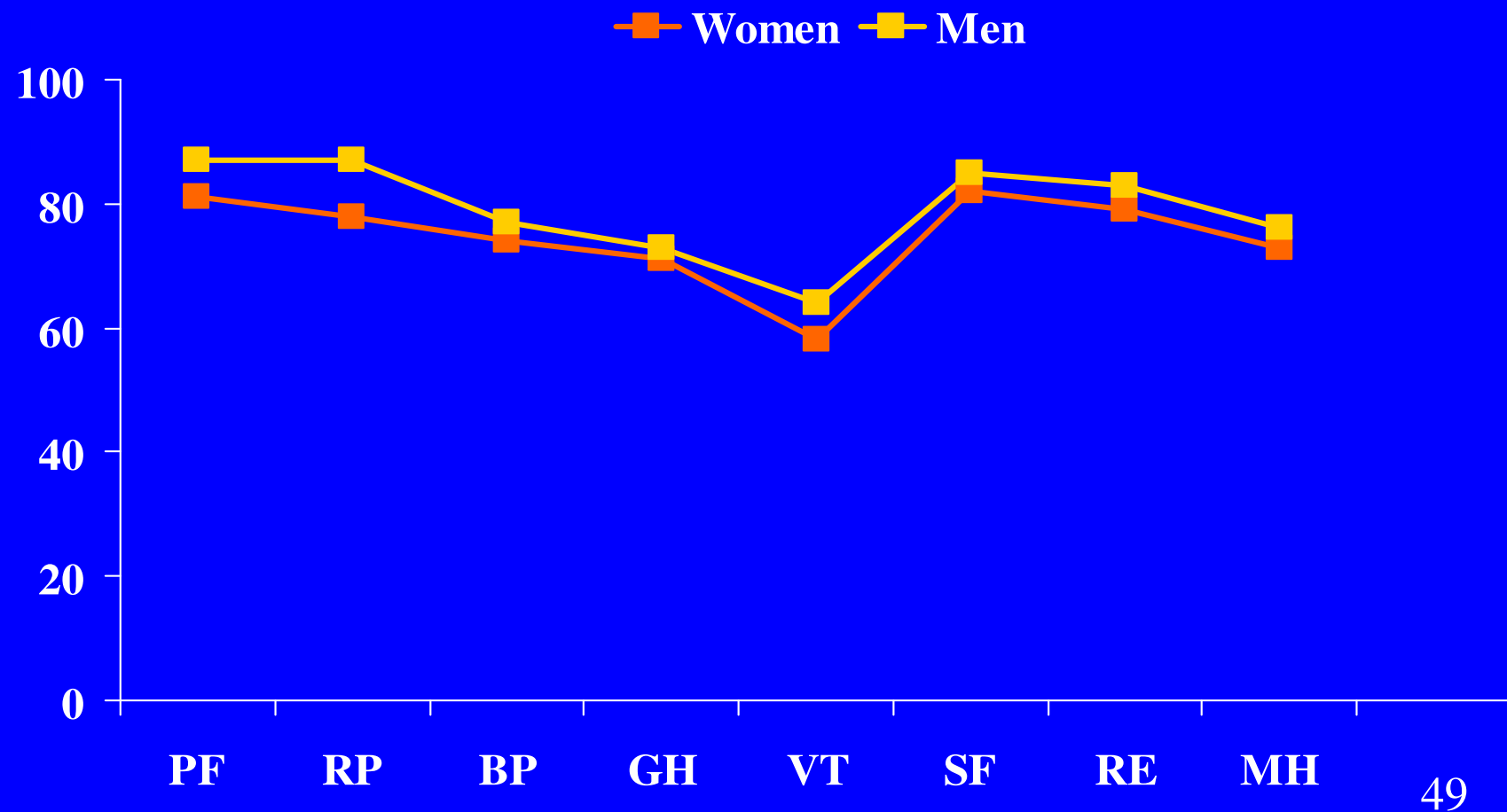
Physical v Mental Health Conditions



Effect of Age on HRQOL



Effect of Gender on HRQOL



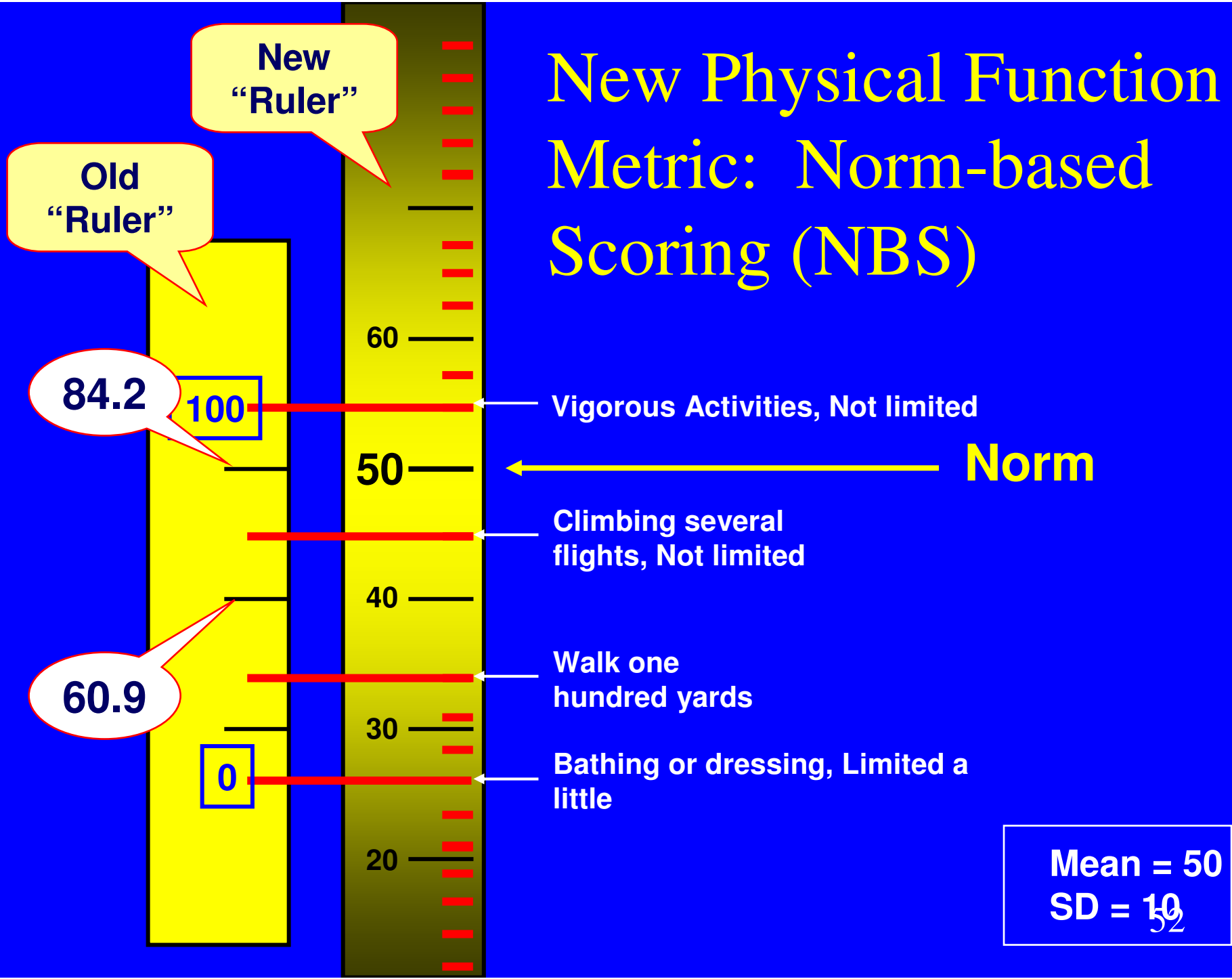
Don't be confounded by age and sex!

- Because mean HRQOL scores vary systematically with age and gender, valid comparisons of sample data with norms and reference values require suitable adjustment
 - Direct or indirect standardisation - epidemiology
 - See Hjerstad et al, European Journal of Cancer 1998

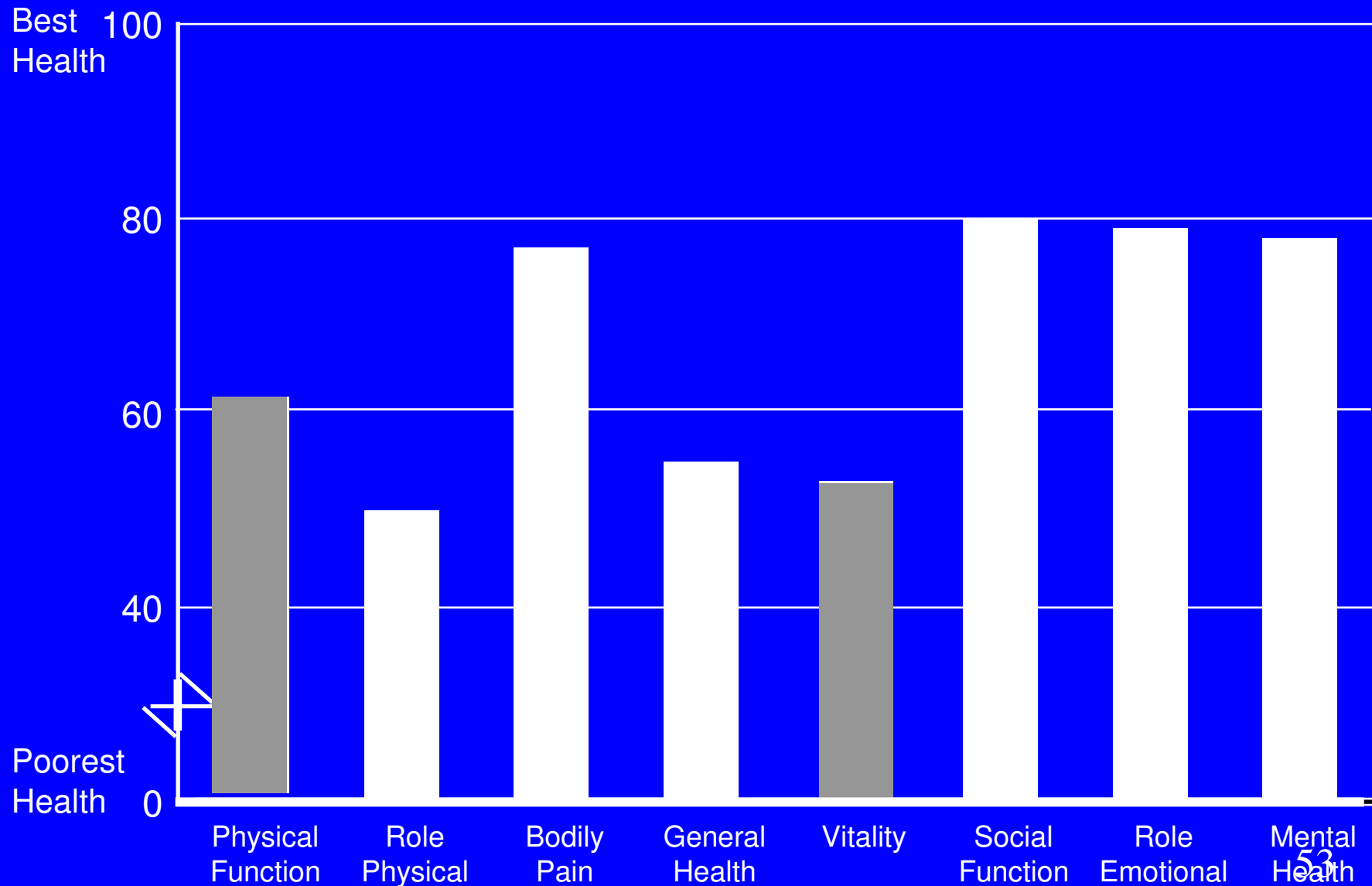
Norm-based scoring

- SF-36 used to be scored 0-100 range
 - linear transformation of raw scores
 - Problems
 - ceiling and floor effects
 - Scores are not really comparable across dimensions
- Now moving to “norm-based” scoring
 - Mean of 50 and SD of 10
 - Normed against the general population
 - Note: for Australian data, need to norm to Aust gen pop values
- Lots of info and explanations at the website
 - <http://www.sf-36.org>

New Physical Function Metric: Norm-based Scoring (NBS)

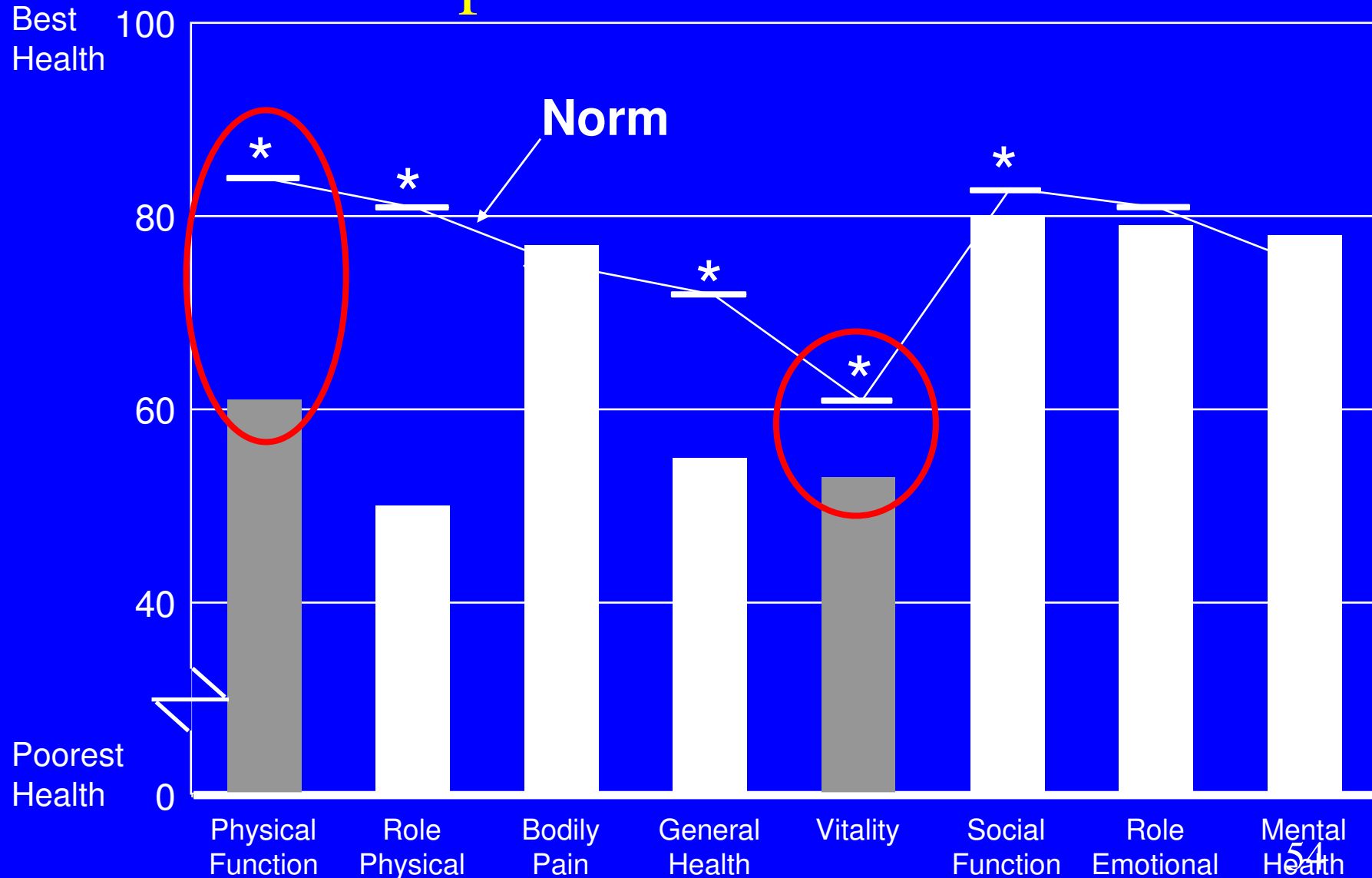


SF-36 Health Profile for Adults with Asthma



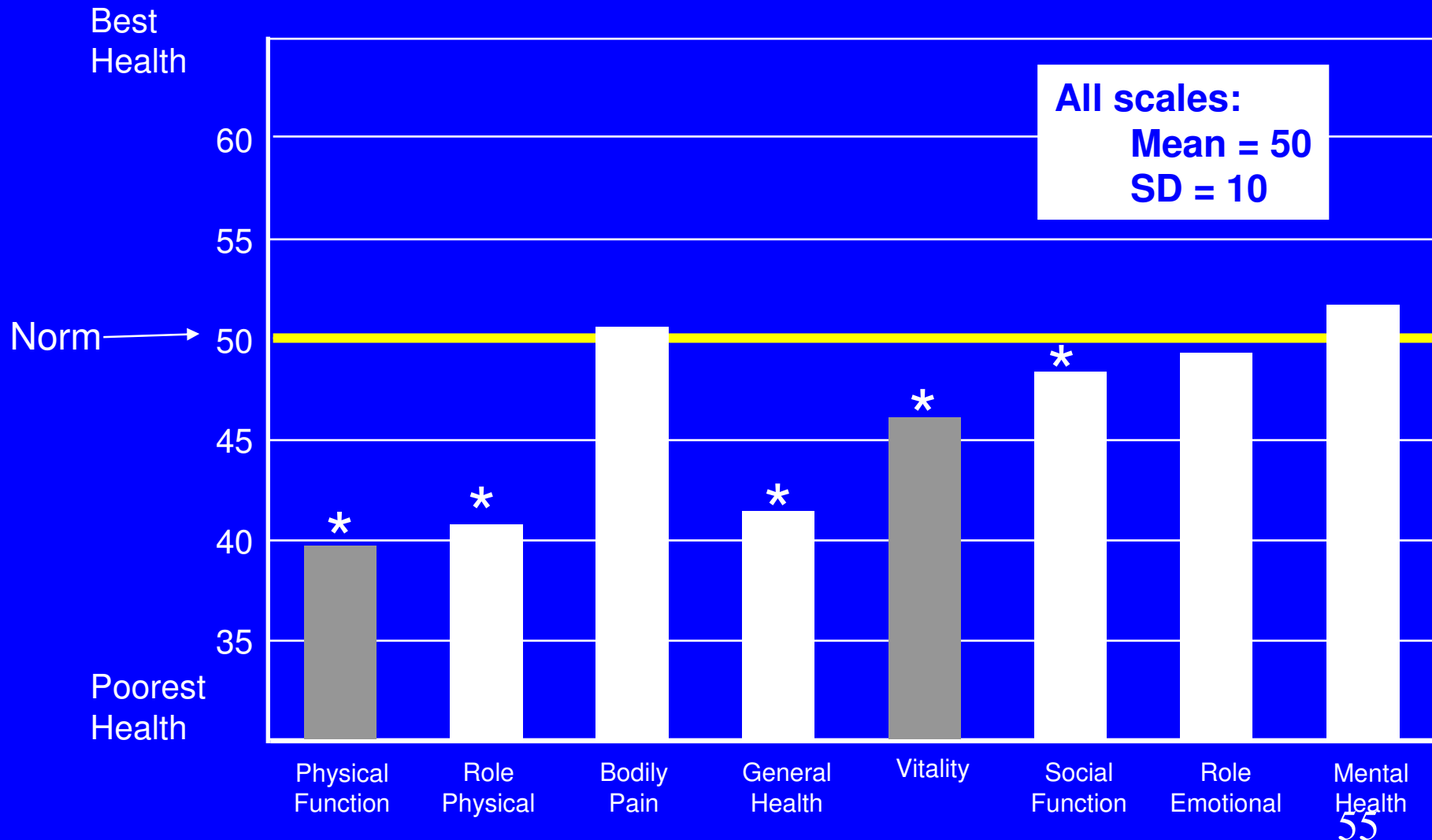
Source: Okamoto 1996

SF-36 Health Profile: Adults with Asthma Compared with U.S. Norm



Source: Okamoto, 1996, * Norm significantly higher

Norm-based Scoring of SF-36 Profile, Adults with Asthma



Adapted from: Okamoto, 1996

Strengths & Weaknesses of Norms and Reference Values

- scores on a particular HRQOL instrument may be compared to norms or reference values for that instrument to give a sense of the impact of disease or treatment on HRQOL
 - can be used for to interpret either group-based or individual score
- BUT
 - not all instruments have norms & reference values
 - but don't give much of a sense of clinical relevance of deviations from “typical” scores

Externally-referenced approaches

2. Anchor-based distributions

- Express differences in HRQOL scores in terms of other more familiar variables
 - symptoms
 - disease severity
 - health care utilization
 - life events - job loss, death of spouse
- Requirements:
 - anchor itself must be interpretable
 - theoretical basis for relationship between anchor & HRQOL domain(s)
 - hence, empirical correlation between anchor and HRQOL domain(s)

Clinical anchors = “known groups”

- Link established classification systems, with well-understood clinical relevance, to novel QOL scales
- Often used to test “clinical” criterion validity
 - So evidence builds up during validation phase

e.g., cancer & ECOG performance status

- 0 = asymptomatic
- 1 = symptomatic but not much impact on daily function
- 2 = significant impact on daily function
- 3 = confined to bed or chair <50% of the day
- 4 = confined to bed or chair >50% of the day

E.g., ECOG & cancer-specific QLQ-C30 (scaled 0-100)

King 1996

ECOG composition of the group					Study #	N	Mean QOL scores		
0	1	2	3	4			Global	Physical	Emotional
100%					16	256	68	85	73
78%	22%				14	68	65	79	73
29%	71%				6	218	61	73	71
28%	72%				8	344	58	66	73
*	*				13	23	57	67	73
	100%				16	226	54	68	66
		100%			13	19	55	56	73
		100%			16	52	37	41	61
		78%	21%	1%	6	84	45	47	67
		70%	25%	5%	8	177	42	40	67
		47%	43%	10%	14	21	58	49	80
			*	*	13	20	38	29	6359

E.g., Diagnosis & extent of cancer

- EORTC QLQ-C30 mean Global QOL scores:
 - heterogeneous cancers, NED (85%) 76
 - heterogeneous “ localized disease 64
 - heterogeneous “ metastatic disease 54
 - recurrent high-grade gliomas 60
 - advanced prostate cancer 45
 - lung cancer, >10% weight loss 43
- This sort of evidence accumulates during validation phase and then in subsequent application
- To be useful in interpretation, requires clinical content knowledge, so in turn, is useful to clinicians

Externally-referenced approaches

3. Global ratings of change

- A patient-based anchor
 - Patient's retrospective assessment of global change
- Patients rate own change as worse /better in each domain of interest:
 - Jaeschke et al 1989 - respiratory & heart
 - Juniper et al 1994 - asthma
 - Osoba et al 1998 – cancer
- Used to determine the “minimum clinically important difference (MCID) (MID)”, “subjectively significant difference (SSD)”

Global ratings of change - summary

Osoba

- Data collection:
 - Prospective QLQ-C30 (0-100 scale) + retrospective change (SSD) for Phys. Emot. & Soc. functioning + global QOL
- Subjective Significance Questionnaire (SSQ) -
Response options: Worse / Same / Better
 - Worse: -1 a little worse to -3 a very much worse
 - Better: +1 a little better to +3 a very much better

-3 -2 -1 0 1 2 3
- Classification of change
 - Small -1 or +1; Moderate -2 or +2; Large -3 or +3

Osoba et al (1998) – similar approach: “subjectively significant difference”

- Data collection:
 - Prospective QLQ-C30 (0-100 scale) + retrospective global change (SSD) for physical, emotional & social functioning
- Results:
- Retro-change score Prospective change score
 - ‘no change’ = less than 5 units change
 - ‘a little’ = 5-10 units change
 - ‘moderate’ = 10-20 units change
 - ‘very much’ = +20 units change

Issues with retrospective global assessment of change as an anchor for interpretation

- individual patient level versus group-based average?
 - Large variation between individuals – is 0.5 the threshold for everyone?
- Retrospective – may be prone to recall bias and response shift (adaptation to change, internal recalibration)
- Single item – does not have psychometric rigour of multi-item scales

Minimum “clinically” important difference?

- Ongoing debate about terminology
 - Does retrospective global assessment really give MCID?
 - Or is it minimum “discernible” difference? Norman 2003
- MCID remains elusive
 - May differ with perspective & context
 - Individual v group
 - Clinician v patient
 - Improvement v deterioration
 - Very well v very sick
 - Physical v emotional function
 - Old v young

Summary of empirical evidence rescaled 0-100

	Small*	Medium	Large
Jaeschke 1989	8	15	20
Juniper 1994	8	17	25
Redelmeier 1996	8		
Osoba 1998	5-10	10-20	20+
King 1996, 2001	2-5	10-15	20

* small (relative to med & large) but not necessarily MCID

Today's best guesstimate

- It seems most evidence & methods are leading to the same/similar result
- A priori, a shift of 1/2 SD on any domain or individual item is likely to be clinically significant
- protocol-specific exceptions can be defined a priori with supportive evidence
 - e.g. 1/4 S.D. is important here because...

Review of externally-referenced approaches to interpretation

Compare HRQOL data with info additional to HRQOL data in the primary data set	Individuals		Groups	
	Cross-sectional	Longitudinal	Cross-sectional	Longitudinal
Clinical known-groups			✓	✓
Norms & reference values	✓	✓	✓	✓
Global ratings of change		✓	✓	✓

✓ approach is useful

✗ approach is feasible but not appropriate

Blank cells indicate approach is not feasible, appropriate or useful

A common sense, context-specific approach to MCID when planning a study

- Think about the context – disease & treatment
- Discuss with clinicians what effects they expect the treatment will have on HRQOL
- Choose a HR-QOL questionnaire that will detect those effects
- Mock survey exercise- fill out as if pre/post tx or comparative groups

Mock exercise - *a priori*

- Consider effect of treatment on HRQOL in very specific terms:
 - Which items do the clinicians expect will be affected?
 - What proportion of patients need to be affected by how much (i.e., how many points on the response scale for each target item?) to be considered a clinically important effect?
 - Calculate the projected mean difference, and use as MCID for sample size calculations

Missing Data

Fairclough 2002

- Incomplete data make interpretation problematic
- Why? Because missing data tends to be “informative” or “non-ignorable”; it tends to be associated with sicker people/occasions
- Avoid missing data as much as possible
- Account for missing data that was unavoidable

Clinical Example

Trastuzumab and chemotherapy versus chemotherapy alone in HER2/neu-overexpressing, metastatic breast cancer

- Slamon D et al, N Engl J Med 2001;344:783-92
- Osoba D et al, J Clin Oncol 2002;20:3106-13

General approach: “Analysis and interpretation of HRQOL data from clinical trials: basic approach of the National Cancer Institute Canada Clinical Trials Group”

- Osoba D et al, Eur J Cancer 2005; 41: 280-287

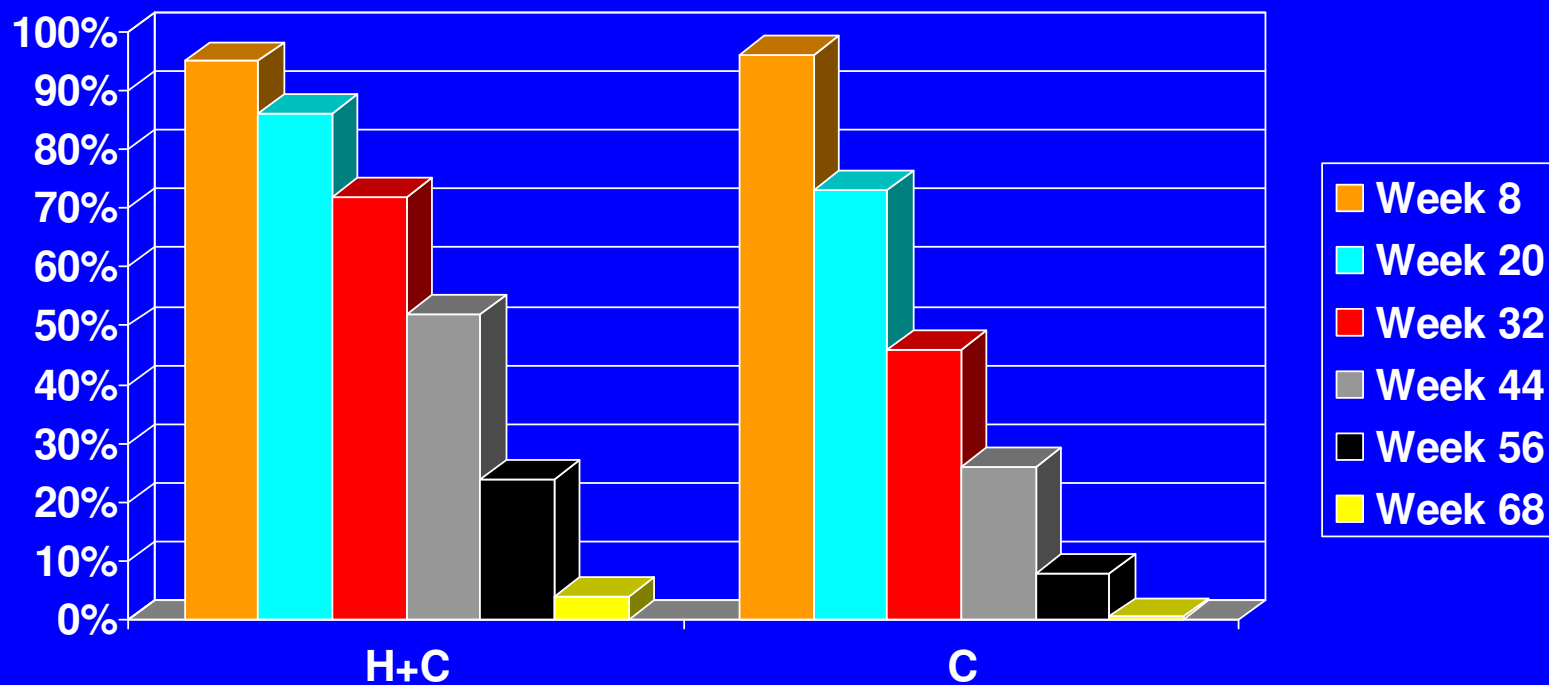
Trastuzumab and chemotherapy

- 400 women with metastatic breast cancer were randomized to receive either trastuzumab and chemotherapy or chemotherapy alone
- global QOL, physical, role, social and emotional functioning and fatigue measured with EORTC QLQ-C30

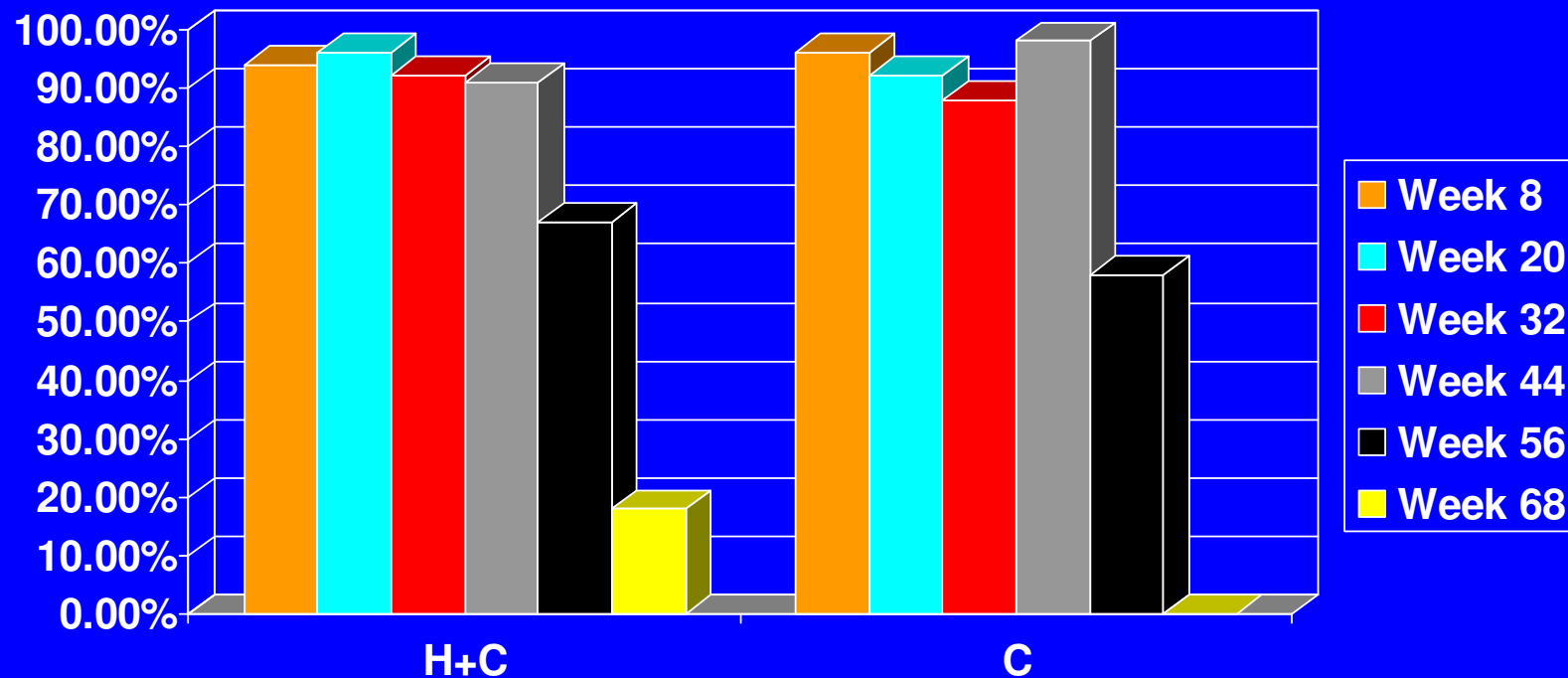
Analysis Approach

- are baseline demographics and HRQOL scores balanced between arms?
- show attrition in number of patients and/or number of completed questionnaires over time

Proportions of Received HRQOL Completions / Baseline Completions

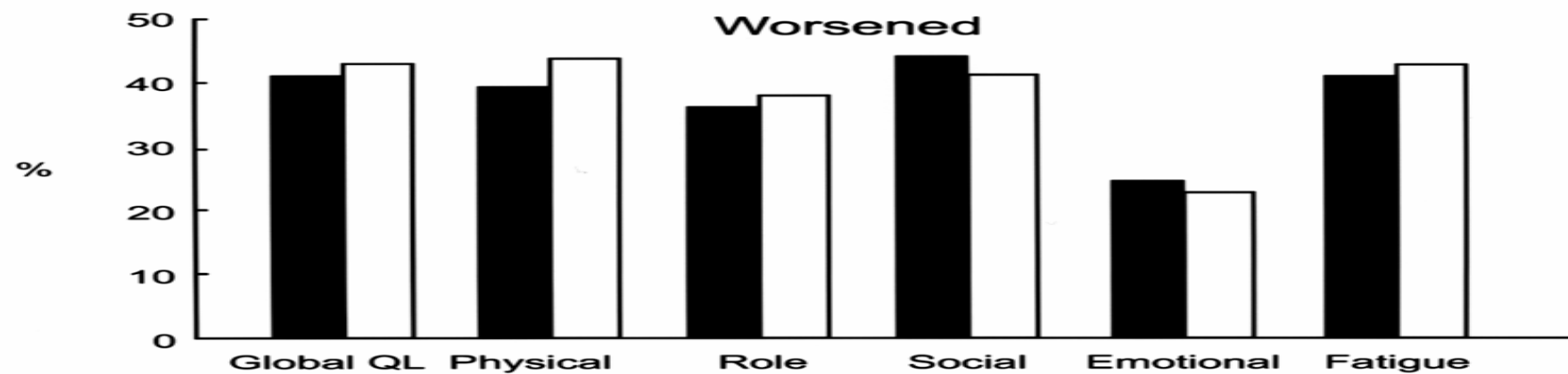
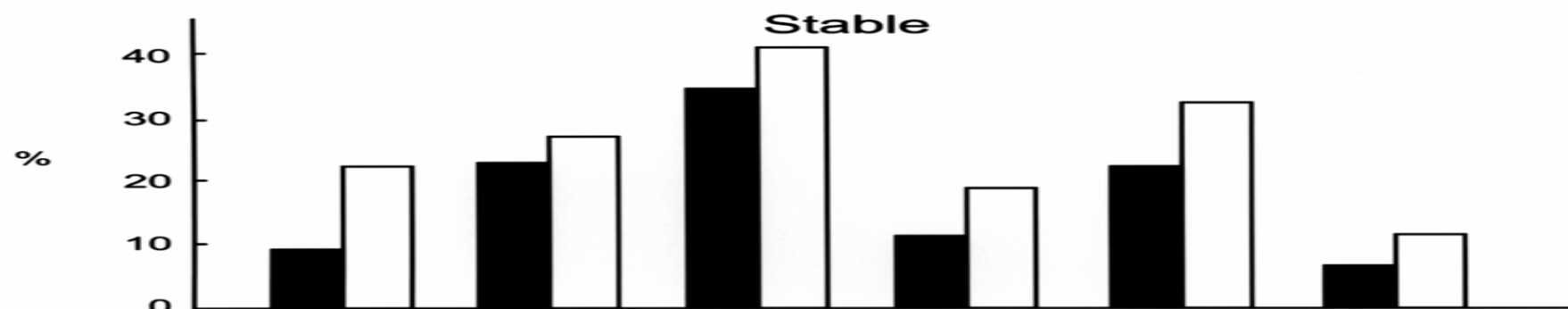
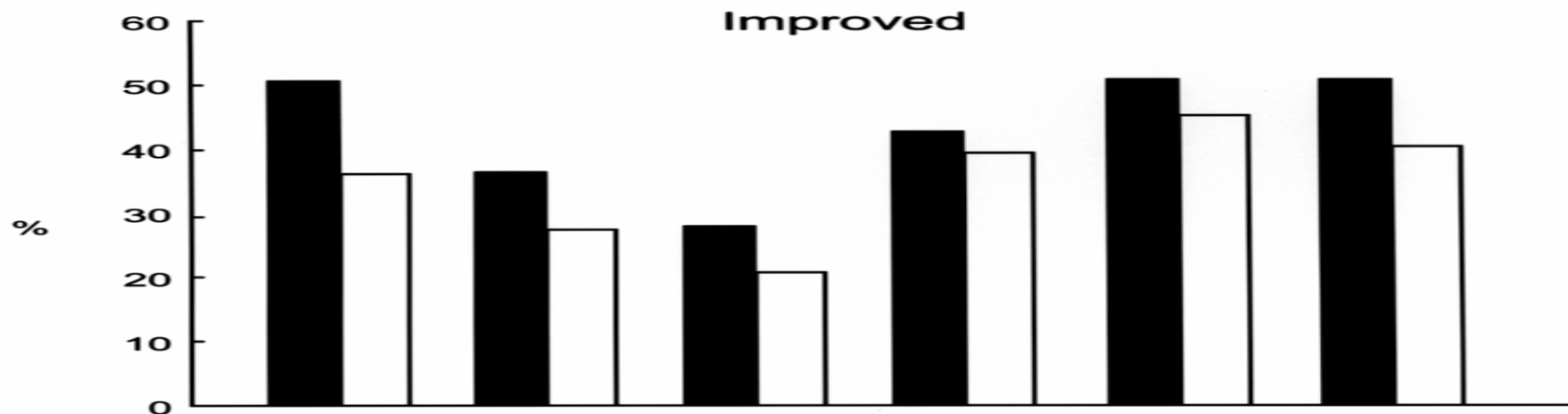


Proportions of Received HRQOL Completions/Expected Completions



Analysis Approach - cont'd

- calculate mean change scores from baseline for each on-study time point
- determine if the change scores are statistically significantly different from baseline (within and between treatment groups)
- provide effect sizes for the observed differences
- show proportions of patients who improve/deteriorate in each arm by preset cut-point e.g., $>/+ 10\%$



Domains

Trastuzumab plus chemotherapy

Chemotherapy

Summary

- various approaches to interpretation
 - each has strengths & weaknesses
 - appropriate / feasible in different circumstances
 - may be used to complement each other

Questions?

Thank you