Challenges in evaluating PROM scores

MATS LUNDSTRÖM

Background

Using a PROM in daily practice is a challenge per se

- Time consuming
- Expensive
- Logistics

But we think it is justified

- Because the patient’s perspective on a disease and/or a treatment is important
- This information may change our treatment policy or may result in clinical improvement work
- It may also change our indications for a treatment
- It can help patients in decision about a treatment

Demands for an effective use of a PROM

Choosing the right PROM or PROMS for a condition

- Measuring one or more dimensions?

Understanding the results – scoring

- Norm data, MID?

MID – minimal important difference

The minimal important difference refers to

”the smallest amount of benefit a patient can recognize and value”

MID, Minimal Important Difference

Recognized improvement:

$\text{PROM}_{\text{outcome}} - \text{PROM}_{\text{baseline}} > \text{MID}$

No recognized improvement:

$\text{PROM}_{\text{outcome}} - \text{PROM}_{\text{baseline}} \leq \text{MID}$

MID: Anchor based

Conditions:

PROM outcome, PROM baseline;

A relevant anchor-question at follow up use to be of Likert type

MID: Mean improvement for patients that recognized a relevant improvement according to the anchor-question

MID: Based on data from the investigation

Standard Deviation (SD) of PROM $_{\text{(baseline)}}$

”Standard” effect size (ES): here Cohen's 0.2, 0.5 or 0.8

MID = SD(PROM $_{\text{(baseline)}}$) X ES

Example: Catquest-9SF 2015; Std. = 2.09

2.09 x 0.8 = 1.67 (MID)

The change in score varied from -12 to + 7.

1.4% had a decrease more than MID. 26.2% had a change within MID. 72.4% had an improvement over MID

MID conditions - 1

Respondents with a high baseline score cannot achieve a relevant improvement:

$PROM_{max} – PROM_{baseline} \leq MID$

1.9% in the Catquest-9SF database had such a high preoperative score that they cannot increase with 1.67 or more.

MID conditions - 2

MID is the same for all:

This can be questioned because of regression to mean:

It is probable that respondents with worse $\text{PROM}_{\text{baseline}}$

Improves more than respondents with better $\text{PROM}_{\text{baseline}}$

Recent publications

A comparison between different outcome measures based on ‘‘meaningful important differences’’ in patients with lumbar spinal stenosis

Maria M. Wertli[1,2,3]
- Franziska Christina Buletti[1]
- Ulrike Held[1]
-

Eva Rasmussen-Barr[2,4]
- Sherri Weiser[2]
- Jakob M. Burgstaller[1]
- Johann Steurer[1]

Eur Spine J

DOI 10.1007/s00586-016-4587-0

“One would expect that two instruments that are valid to measure pain would be similarly sensitive to change and a high agreement between the proportions of patients with MID can be found for instruments that measure the same domain.”

Wertli et al. 2016: 466 patients completed a baseline and 6 months follow up assessment

3 PAIN MEASURES, ALL RECOGNIZED AND VALIDATED (CLASSICAL TEST THEORY)

SSM (Spinal Stenosis Measure) Sy

NRS (Numeric Rating Scale)

FT (Feeling thermometer)

MID change varied between 40% and 65% depending on outcome measures and cut-offs

3 FUNCTION MEASURES, ALL RECOGNIZED AND VALIDATED (CLASSICAL TEST THEORY)

SSM F

RMQ (Roland Morris Questionnaire)

FT

MID change varied between 40% and 70% depending on outcome measures and cut-offs

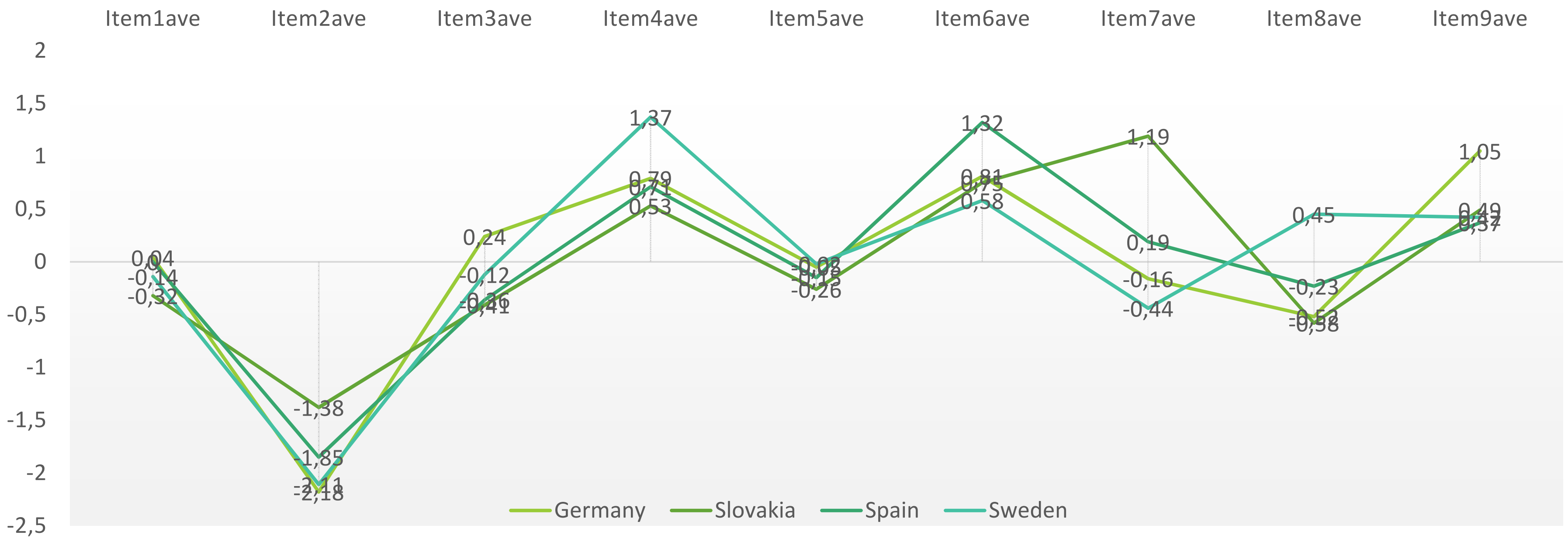
Another finding was that the MID change in pain measure and function measure disagreed in about 30% of the cases. So measuring different domains may give different outcomes.

Scoring algorithm in different language versions of a questionnaire

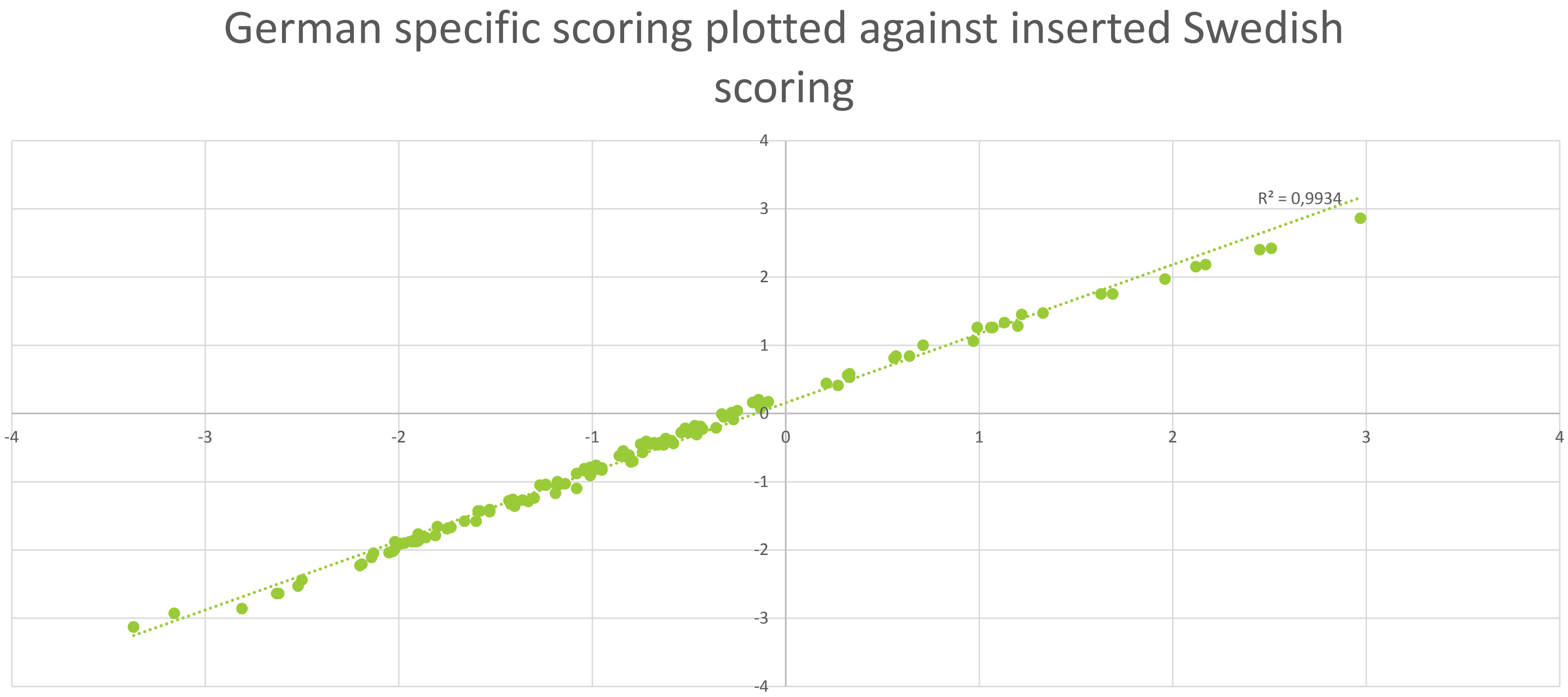
Country	Number	Mean age	Min	Max	% female
Germany	133	73.6	41	92	63.2
Slovakia	248	69.3	38	90	58.6
Spain	294	61.7	25	93	61.9
Sweden	295	74	50	91	61.7

Comparison of item scores in 4 different language populations

Average item Rasch score, Item 1-9, 4 languages

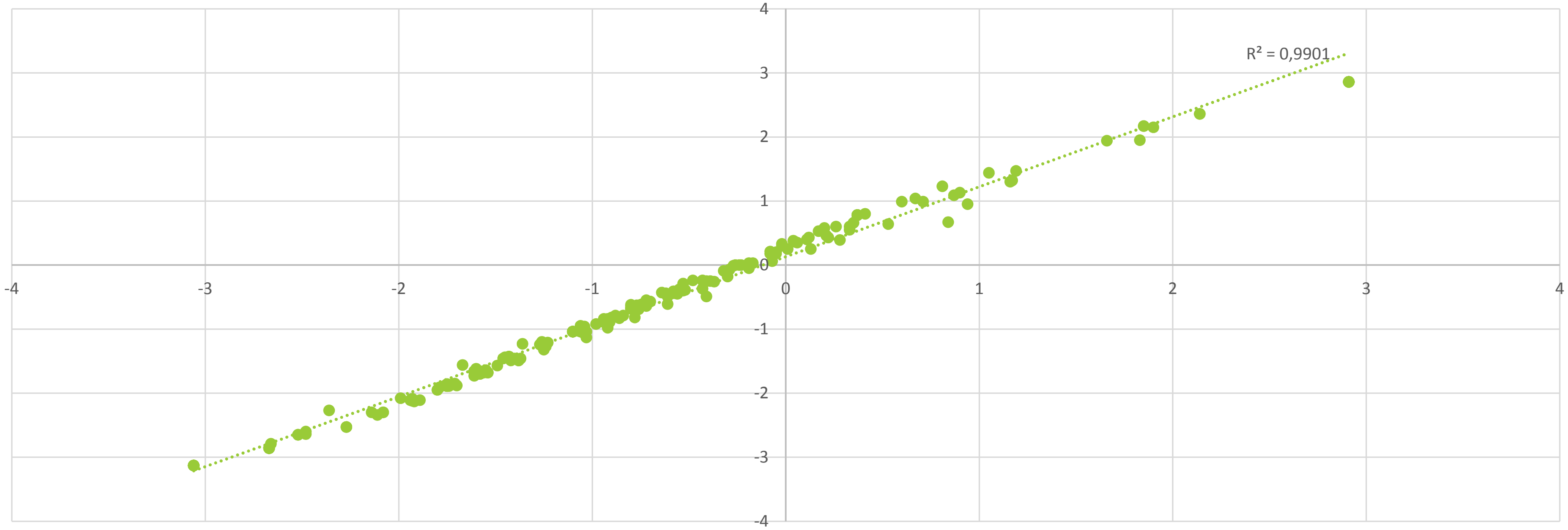


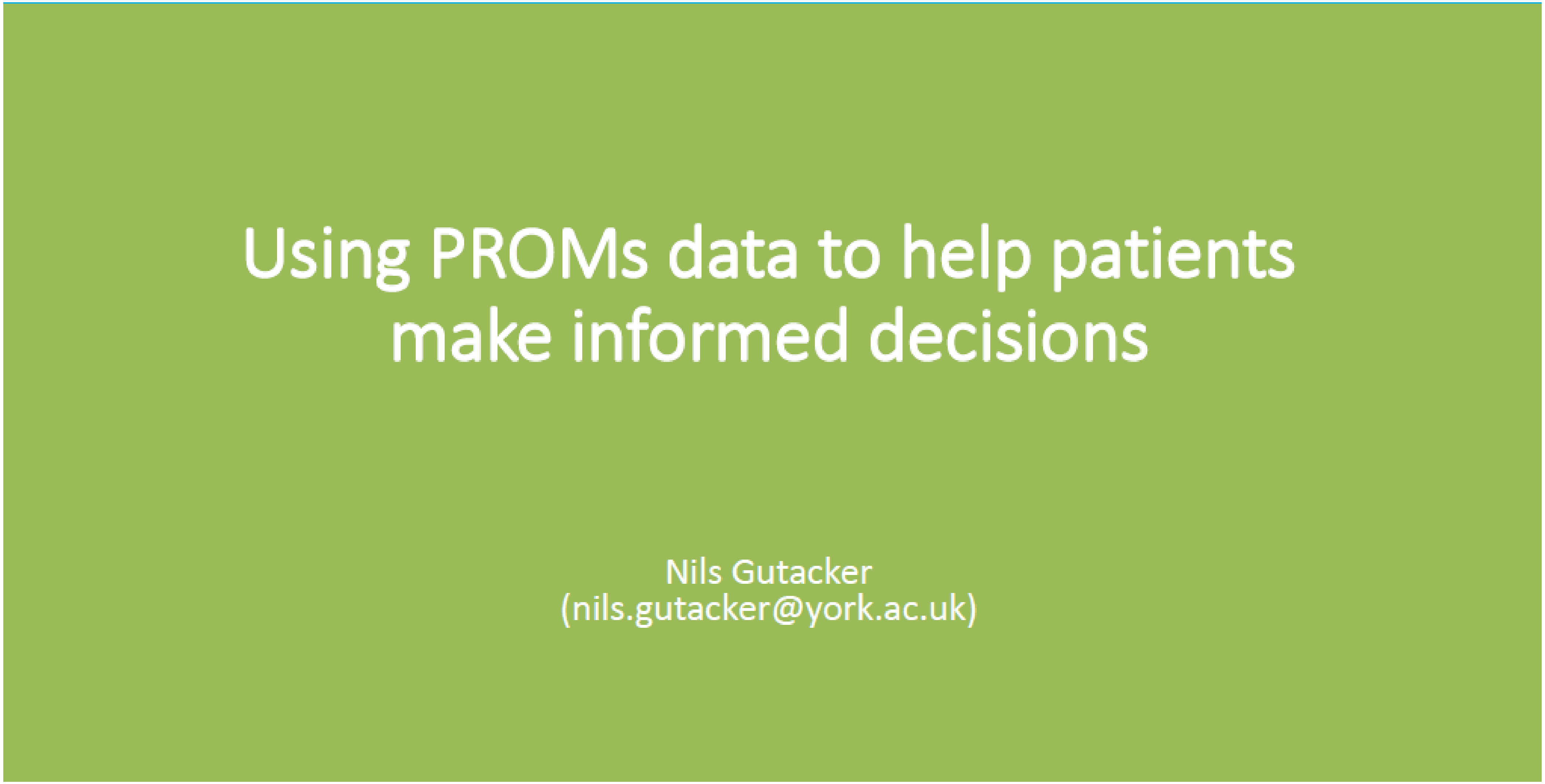
German scoring vs Swedish scoring



Slovakian scoring vs Swedish scoring

Slovakian specific scoring plotted against inserted Swedish scoring





Using PROMs data to help patients make informed decisions

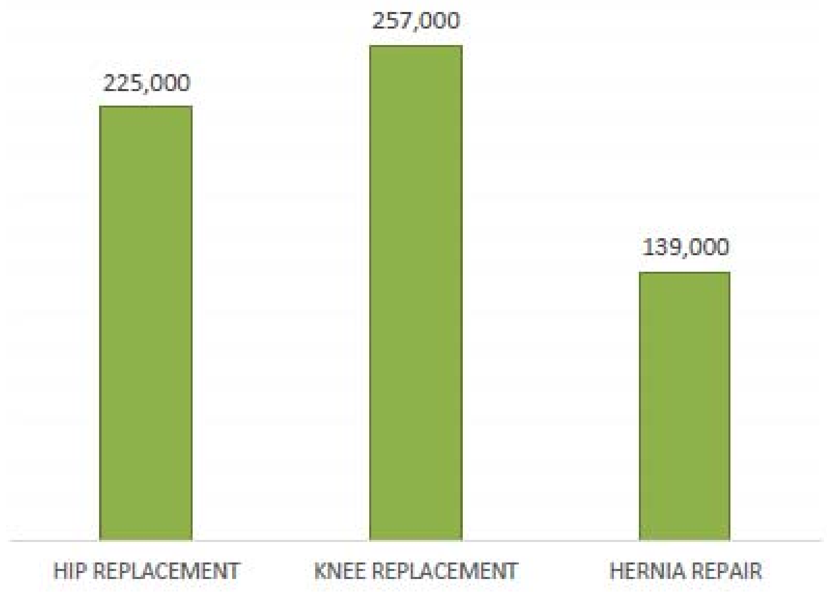
Nils Gutacker
(nils.gutacker@york.ac.uk)

Courtesy Dr. Nils Gutacker

English PROM programme

- Started April 2009
- Collects HRQoL data
  + basic demographics
  before and after surgery
- All NHS patients undergoing
  hip/knee/hernia surgery

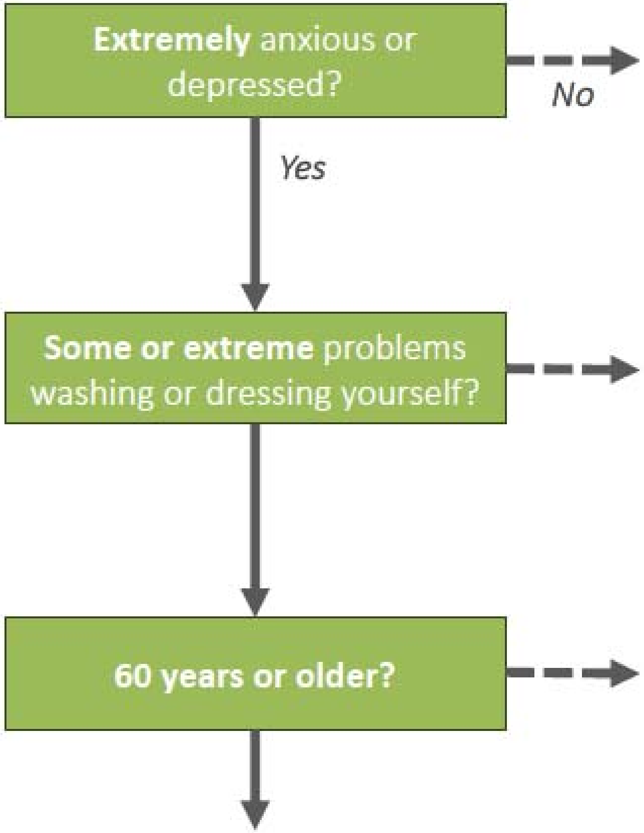
Complete pre- and post-operative PROMs collected 2009/10 to 2013/14



Courtesy Dr. Nils Gutacker

Developing a tool to inform patients

- Uses anonymised EQ-5D data
- Regression tree analysis (CART) groups patients with similar post-operative health levels
- Based on 8 factors: age, sex, symptoms, five HRQoL items



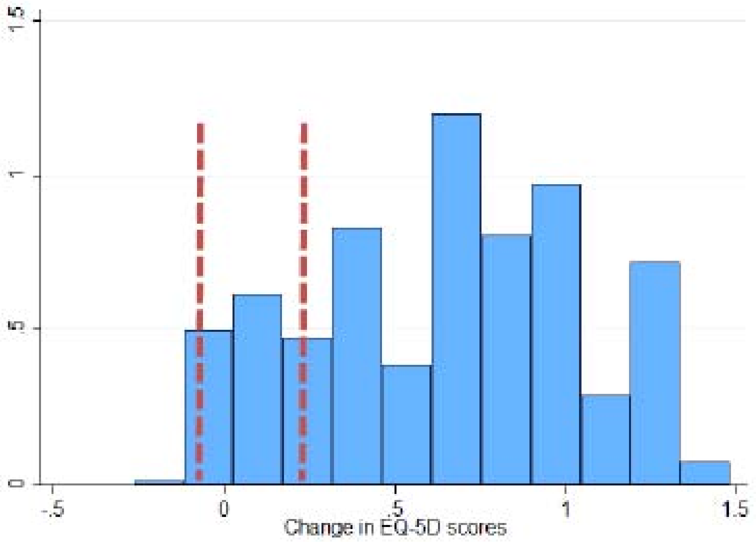
Extremely anxious or depressed?

Some or extreme problems washing or dressing yourself?

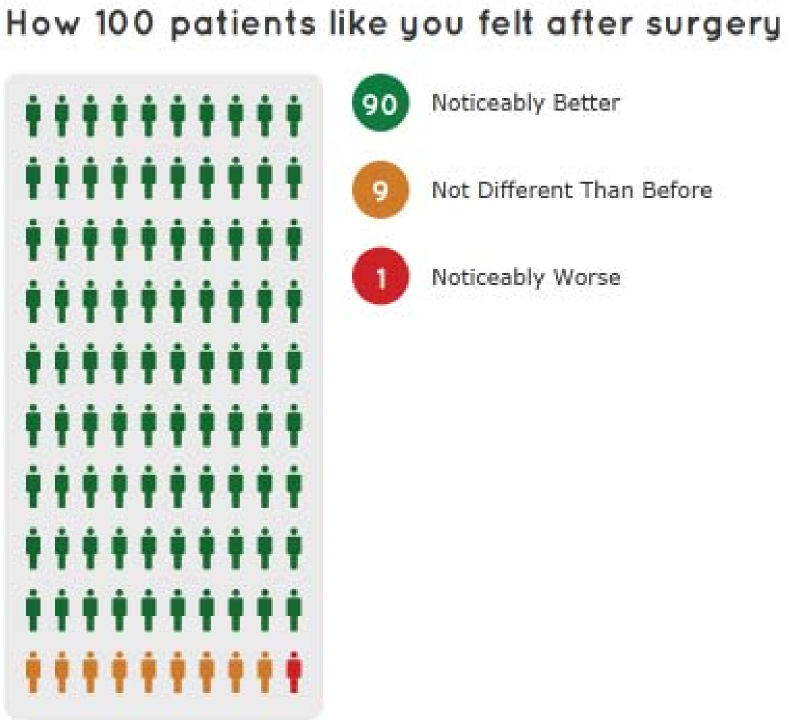
60 years or older?

Courtesy Dr. Nils Gutacker

Making the data meaningful to non-technical audiences

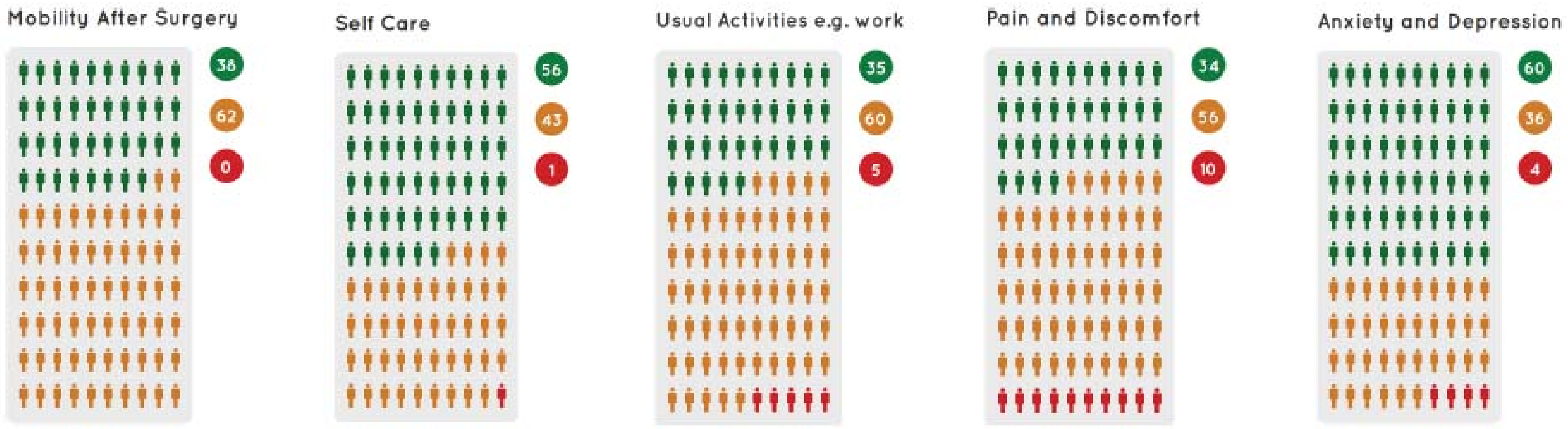


Minimally important difference (MID)


How 100 patients like you felt after surgery

Courtesy Dr. Nils Gutacker

Additional detail by HRQoL domain



Courtesy Dr. Nils Gutacker

Predict meaningful improvement by use of PROM score

Berliner JL, Brodke DJ, Chan V, SooHoo NF, Bozic KJ.

Can preoperative patient-reported measures be used to predict meaningful improvement in function after TKA?

Clin Orthop Relat Res. 2016 Mar 8 [Epub ahead of print]

Study outline

562 patients going through primary unilateral TKA

Patients completed 2 PROMs before and 1 year after surgery

- Knee injury and Osteoarthritis Outcome Score (KOOS)
- SF-12 v. 2 (SF12v2)

Minimum clinical important differences (MCIDs) were calculated with a distribution-based method to define meaningful clinical improvement

Physical component summary scores were calculated

Threshold values for preoperative KOOS and SF12v2 scores were determined. Threshold values defined the point after which the likelihood o clinically meaningful improvement began to diminish

Results

Patients scoring above thresholds indicating a better preoperative function were less likely to experience a clinically meaningful improvement in function after TKA.

When accounting for mental and emotional health the predictive ability improved.

Patients with worse baseline mental and emotional health had a lower probability of experiencing clinically meaningful improvement after TKA