



MEDIZINISCHE UNIVERSITÄT
INNSBRUCK

Die Entwicklung und Verwendung von Schwellenwerten für klinische Relevanz zur Unterstützung der Interpretation von patientenberichteten Endpunkten

Johannes M. Giesinger

Univ.-Klinik für Psychiatrie II

Medizinische Universität Innsbruck

Introduction

- Patient-reported outcome (PRO) measures are increasingly used to assess endpoints in clinical trials
- Routine PRO monitoring in daily practice has demonstrated important clinical benefits including
 - Higher patient satisfaction (e.g. Basch et al. 2016, Kotronoulas et al. 2014)
 - Better symptom management (e.g. Kotronoulas et al. 2014)
 - Improved survival rates (e.g. Basch et al. 2017; Denis et al. 2017, 2019)
- **PRO measures are useful, but interpretation of results is challenging as score points are rather abstract!**

Introduction

- Different approaches to facilitating interpretation of scores have been introduced, e.g.:
 - Minimal Important Differences (MID)
 - Normative data (e.g. from general population)
 - Thresholds (cut-off scores) with specific meanings
- But still only few studies (e.g. Snyder 2013, 2015; Jansen 2016) on thresholds supporting interpretation of **absolute scores from an individual patient at a single time point**
→ different from minimal important differences/changes
- Thresholds allow to assess the clinical relevance of scores and to calculate symptom prevalence based on (metric/ordinal) PRO measures

Example for Threshold Development: EORTC QLQ-C30

Development of thresholds for clinical importance for 14 of the 15 scales (excl. global QOL) included in the QLQ-C30:

- A) Evaluation of **what makes a symptom relevant** for patient-clinician consultation based on a mixed methods study with patients and health professionals
- B) Definition of **anchor items** as external criteria for determining thresholds by an expert board based on the information collected in aim A
- C) Determination of **thresholds** for the QLQ-C30 using the anchor items determined in aim B

Cut-off scores should identify symptoms/problems that need to be discussed in the clinical encounter

Collaborators on EORTC QLQ-C30 threshold development

- **Juan I. Arraras** (Hospital of Navarre, Pamplona, Spain)
- **Giovanni Caocci** (University of Cagliari, Italy)
- **Fabio Efficace** (Italian Group for Adult Hematologic Diseases (GIMEMA) Data Center, Rome, Italy)
- **Mogens Groenvold, Morten Petersen** (Bispebjerg Hospital, Copenhagen, Denmark)
- **Bernhard Holzner, Fanny Loth** (Medical University of Innsbruck, Austria)
- **Neil Aaronson, Marieke van Leeuwen** (The Netherlands Cancer Institute, Amsterdam, The Netherlands)
- **John Ramage** (Hampshire Hospitals NHS Foundation Trust, Basingstoke, UK)
- **Krzysztof A. Tomaszewski** (Andrzej Frycz Modrzewski Krakow University, Krakow, Poland)
- **Teresa Young** (Mount Vernon Cancer Centre, Northwood, UK)

Aim A: What makes a symptom relevant for patient-clinician consultation?

- Mixed methods study in cancer patients and health care professionals
- Participants from Austria, Italy, the Netherlands, Poland, Spain, and the UK
- Qualitative interview on what makes a symptom/functional health impairment clinically important
- Quantitative importance ratings for a set of pre-defined criteria of clinical importance

Aim A: What makes a symptom relevant for patient-clinician consultation?

Recruitment in Austria, Italy, The Netherlands, Poland, Spain, UK (N=150)

Sample characteristics		Patients N=83	Health prof. N=67
Age	Mean	60.3 yrs	44.2 yrs
Sex	Women	49.4%	65.7%
Diagnosis (most frequent)	Breast cancer	24.1%	-
	Colorectal cancer	19.3%	-
Cancer stage	Stage III or IV	61.5%	
Professional background	Nurses	-	31.3%
	Oncologists	-	29.9%
	Psycho-oncologists	-	16.4%
	Surgeons	-	10.4%
	Other	-	11.9%

Aim A: Results (qualitative interviews)

Categories	Patients	Health care professionals
problems limits everyday life or daily functioning	64.2%	77.6%
if problem causes other problems	58.0%	29.9%
problem has emotional impact	54.3%	64.2%
duration or frequency of problem	50.6%	49.3%
problem changed from normal levels	34.6%	26.9%
problem has impact on family or partner	29.6%	37.3%
help or treatment is needed	29.6%	23.9%
problem might indicate other problem	23.5%	17.9%

Aim A: Results (quantitative ratings)

	Quite/Very important %	
patient's partner/family worried about the symptom/problem	Patients	88%
	HCPs	82%
symptom/problem is a limitation in everyday life	Patients	95%
	HCPs	100%
patient needs help from the medical treatment	Patients	86%
	HCPs	89%
patient needs help from family/friends	Patients	83%
	HCPs	95%
patient worried about the symptom or problem	Patients	83%
	HCPs	88%

Received: 24 February 2017

Revised: 28 July 2017


Accepted: 24 August 2017

DOI: 10.1002/pon.4548

WILEY

PAPER

A cross-cultural convergent parallel mixed methods study of what makes a cancer-related symptom or functional health problem clinically important

Johannes M. Giesinger¹  | Neil K. Aaronson² | Juan I. Arraras³ | Fabio Efficace⁴ | Mogens Groenvold⁵ | Jacobien M. Kieffer² | Fanny L. Loth¹ | Morten Aa. Petersen⁵ | John Ramage⁶ | Krzysztof A. Tomaszewski⁷ | Teresa Young⁸ | Bernhard Holzner¹ |
on behalf of the EORTC Quality of Life Group

Psycho-Oncology 2017

Aim B: Definition of clinical importance

Consensus discussion within an EORTC Quality of Life Group meeting (Oslo) consisting of members of different professional backgrounds (e.g., oncology, surgery, psychology, nursing, statistics)

Examples for considerations on establishing anchors:

- We decided to use the same set of anchors (with domain-specific wording) for all scales to ensure that the meaning of the thresholds was consistent across domains.
- We excluded the 'duration or frequency' and 'change from normal' aspects because the QLQ-C30 measures symptom severity only.

Aim B: Definition of clinical importance

Final criteria for clinical importance of a symptom or functional health impairment:

During the past week:	Not at all	A little	Quite a bit	Very much
Has shortness of breath limited your daily life?	1	2	3	4
Have you needed any help or care because of shortness of breath?	1	2	3	4
Have you had shortness of breath causing you or your family/partner to worry?	1	2	3	4

Classification rule for cases: selecting a **red** category for any of the anchor items makes a symptom/problem **clinically important**

Aim C: Thresholds for the QLQ-C30 and EORTC CAT Core

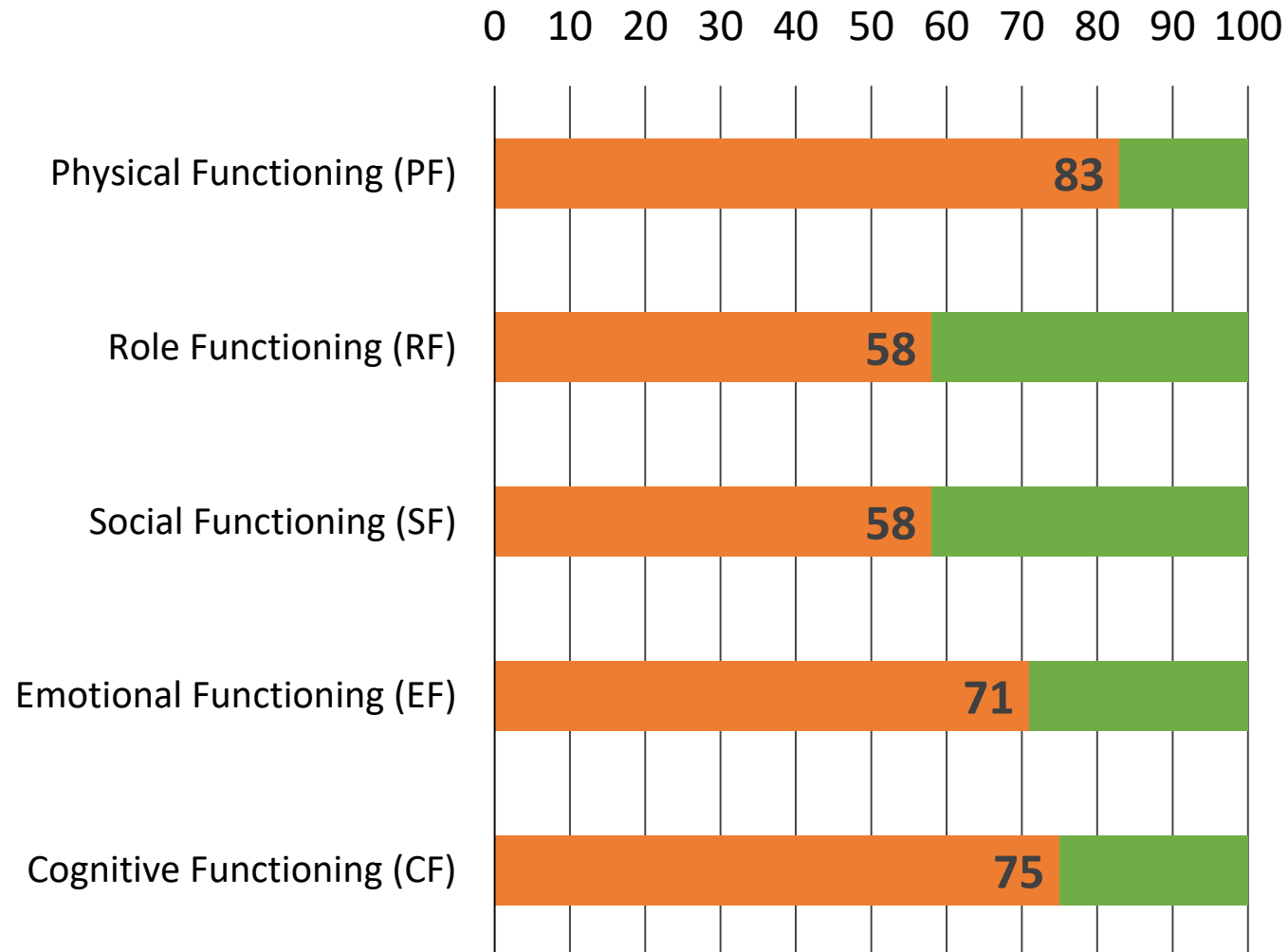
Recruitment in Austria, Italy, The Netherlands, Poland, Spain, UK (N=498)

Sample characteristics		N=498
Age	Mean	60.4 yrs
Sex	Women	55.2%
Diagnosis (most frequent)	Breast cancer	23.6%
	Haematological malign.	13.3%
	Lung cancer	9.9%
Cancer stage	Stage III or IV	57.5%
Treatment status	Current treatment	76.7%
Treatment intention	Curative	59.0%

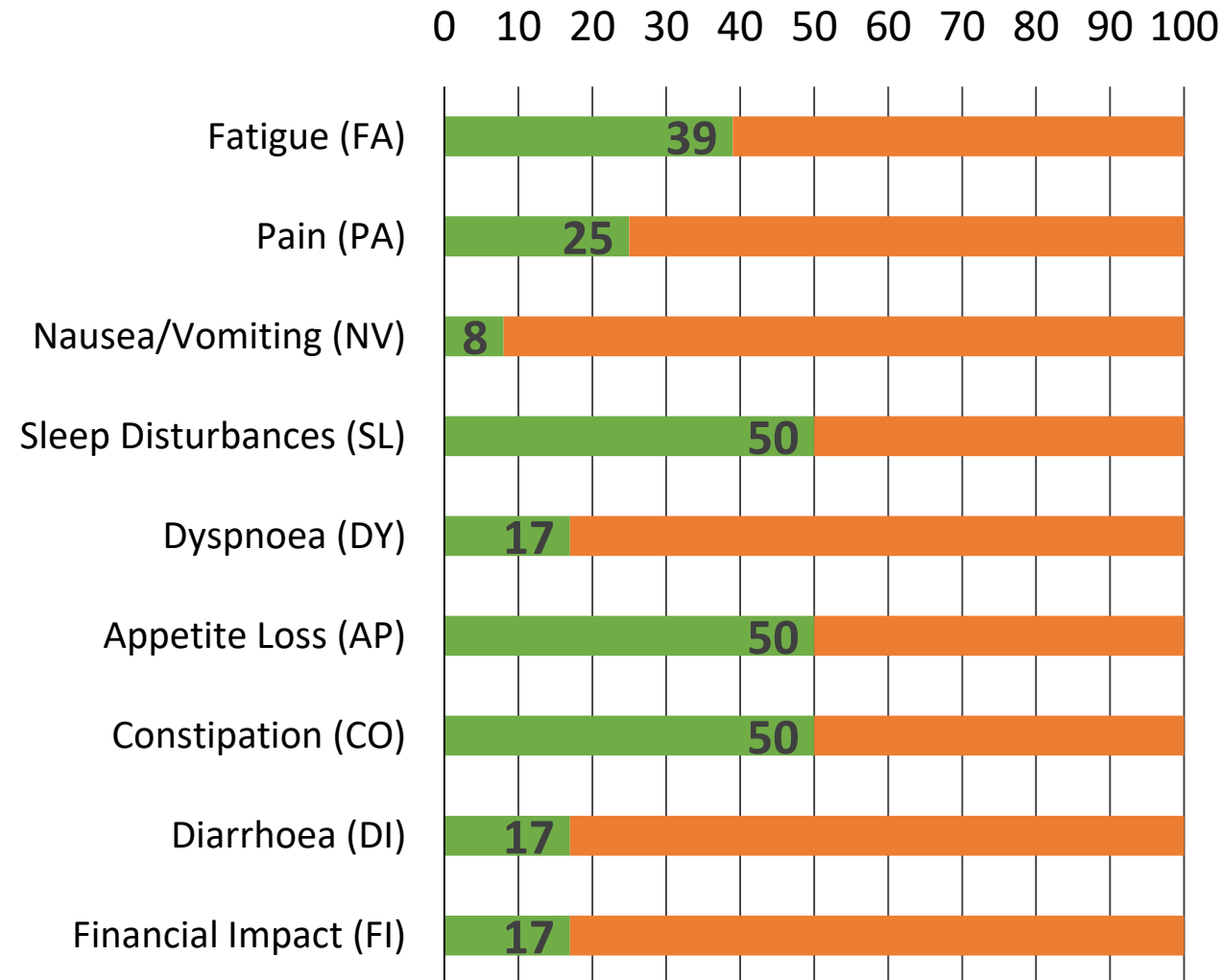
Statistical analysis

- Thresholds were determined using Receiver Operating Characteristic (ROC) analysis:
 - EORTC QLQ-C30 scales were the predictors
 - Binary categories of clinical importance were the criterion
- For each scale, we selected the threshold with the highest Youden's J statistic (the sum of sensitivity and specificity minus one)
- If the Youden's J of two adjacent thresholds differed by less than 0.05, we selected the threshold with higher sensitivity.

Aim C: QLQ-C30 Thresholds for clinical importance



Aim C: QLQ-C30 Thresholds for clinical importance



Aim C: QLQ-C30 Thresholds for clinical importance

Diagnostic characteristics of the thresholds when used to predict clinical importance

QLQ-C30 Functioning	Thresholds	Sensitivity	Specificity	AUC
Physical Functioning (PF)	83	0.79	0.70	0.82
Role Functioning (RF)	58	0.82	0.77	0.87
Social Functioning (SF)	58	0.71	0.77	0.82
Emotional Functioning (EF)	71	0.80	0.76	0.87
Cognitive Functioning (CF)	75	0.77	0.73	0.81

Aim C: QLQ-C30 Thresholds for clinical importance

Diagnostic characteristics of the thresholds when used to predict clinical importance

QLQ-C30 Symptoms	Threshold	Sensitivity	Specificity	AUC
Fatigue (FA)	39	0.89	0.82	0.92
Pain (PA)	25	0.92	0.71	0.90
Nausea/Vomiting (NV)	8	0.86	0.75	0.87
Dyspnoea (DY)	50	0.79	0.84	0.87
Sleep Disturbances (SL)	17	0.97	0.62	0.88
Appetite Loss (AP)	50	0.78	0.90	0.89
Constipation (CO)	50	0.73	0.92	0.86
Diarrhoea (DI)	17	0.88	0.77	0.85
Financial Impact (FI)	17	0.91	0.81	0.91



ORIGINAL ARTICLE

Thresholds for clinical importance were established to improve interpretation of the EORTC QLQ-C30 in clinical practice and research

Johannes M. Giesinger^{a,*}, Fanny L.C. Loth^a, Neil K. Aaronson^b, Juan I. Arraras^c,
Giovanni Caocci^d, Fabio Efficace^e, Mogens Groenvold^f, Marieke van Leeuwen^b,
Morten Aa. Petersen^g, John Ramage^h, Krzysztof A. Tomaszewskiⁱ, Teresa Young^j,
Bernhard Holzner^a, on behalf of the EORTC Quality of Life Group

^aUniversity Hospital of Psychiatry II, Medical University of Innsbruck, Innsbruck, Austria

^bDivision of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute, Amsterdam, the Netherlands

^cMedical Oncology Department, Hospital of Navarre, Pamplona, Spain

^dDepartment of Medical Sciences and Public Health, University of Cagliari, Cagliari, Italy

^eHealth Outcomes Research Unit, Italian Group for Adult Hematologic Diseases (GIMEMA) Data Center, Rome, Italy

^fThe Research Unit, Department of Palliative Medicine, Bispebjerg & Frederiksberg Hospital, University of Copenhagen, Copenhagen, Denmark

^gDepartment of Public Health, University of Copenhagen, Copenhagen, Denmark

Journal of Clinical Epidemiology 2020

Thresholds based on bookmarking

- The PROMIS initiative has developed thresholds based on bookmarking by patients and health care professionals (e.g. Rothrock 2019; Bingham 2021)
- Case vignettes are created based on item text and response categories corresponding to specific scores
- Patients and/or health professionals categorize case vignettes (i.e. symptom/functioning levels) as "within normal limits," "mild", "moderate" or "severe"

Example for boomarking: Rothrock et al. 2019

- Study objective: Establishing severity thresholds for PROMIS measures of physical function, cognitive function, and sleep disturbance in cancer patients
- Sample: 6 patients (focus group), 10 clinicians
- Vignettes were developed with five items.
- Vignette locations were $\frac{1}{2}$ standard deviation (SD) apart
- Range for physical function T=12.5 – 62.5

Case vignettes for bookmarking

Ms. King's Physical Function (T-score=47.5)

- Ms. King had a little difficulty running a short distance (for example, to catch a bus).
- She was somewhat limited in doing vigorous activities, such as running, lifting heavy objects, and participating in strenuous sports.
- She had very little limitation in climbing several flights of stairs and could run errands and shop with no difficulty.
- She had a little difficulty turning faucets on and off.

Mr. Lopez's Physical Function (T-score=52.5)

- Mr. Lopez's was able to run or jog for two miles (3 km) with some difficulty.
- He was not limited at all in hiking a couple of miles on uneven surfaces, including hills, nor was he limited in doing heavy work around the house (like scrubbing floors or lifting or moving heavy furniture).
- He could carry a laundry basket up a flight of stairs without any difficulty.
- He had a little difficulty opening previously opened jars.

T-score: ... 30 35 40 45 50 55 60 65 70 ...

Thresholds based on bookmarking

Established thresholds for PROMIS Physical Function:

	PHYSICAL FUNCTION											
T-Score:	70	65	60	55	50	45	40	35	30	25	20	15
Patient Consensus			WNL*		MILD			MOD			SEVERE	
Clinician Consensus			WNL*		MILD		MOD		SEVERE			

From Rothrock 2019 – Figure 2:

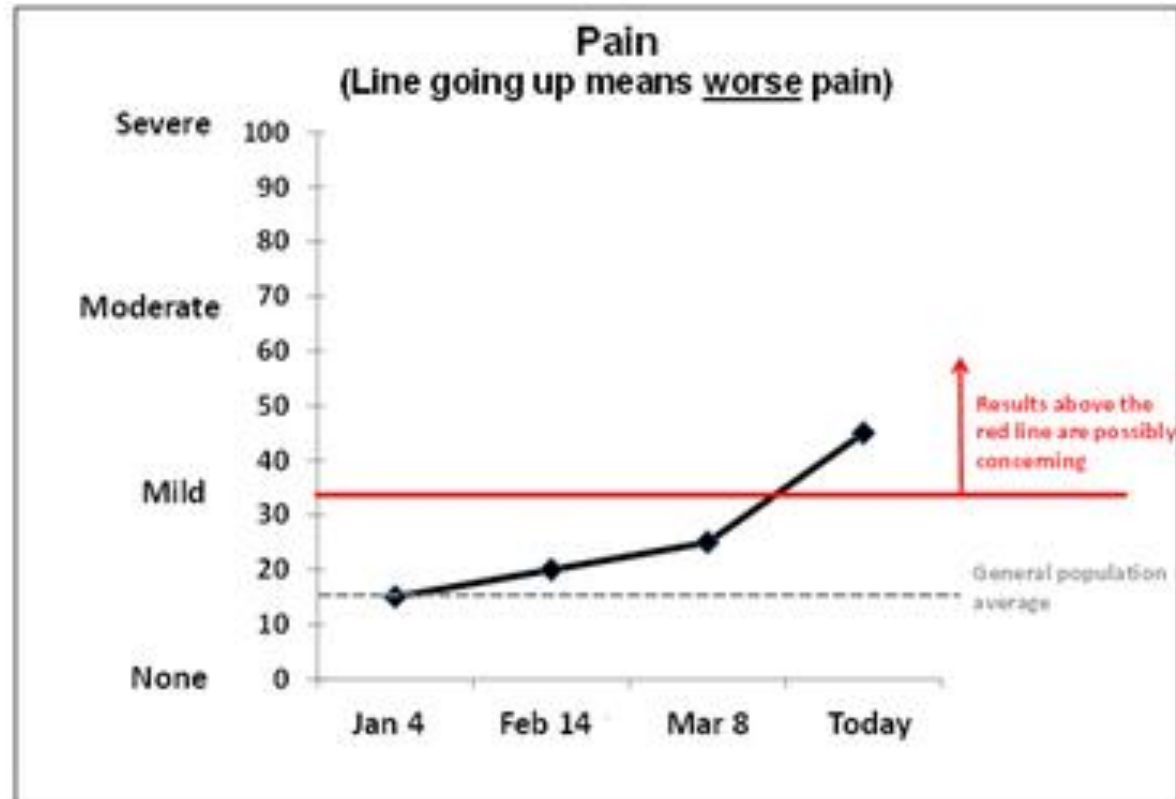
Comparison of Patients' and Clinicians' Thresholds for Levels of Symptoms and Function

* within normal range

(Dis)advantages of bookmarking

- No estimates for sensitivity/specificity, or invariance across patient groups
- Possible thresholds differ by 0.5 standard deviations
- No explicit definition of the meaning of the thresholds
- Small sample size
- Less resources/time required

Thresholds and graphical result presentation



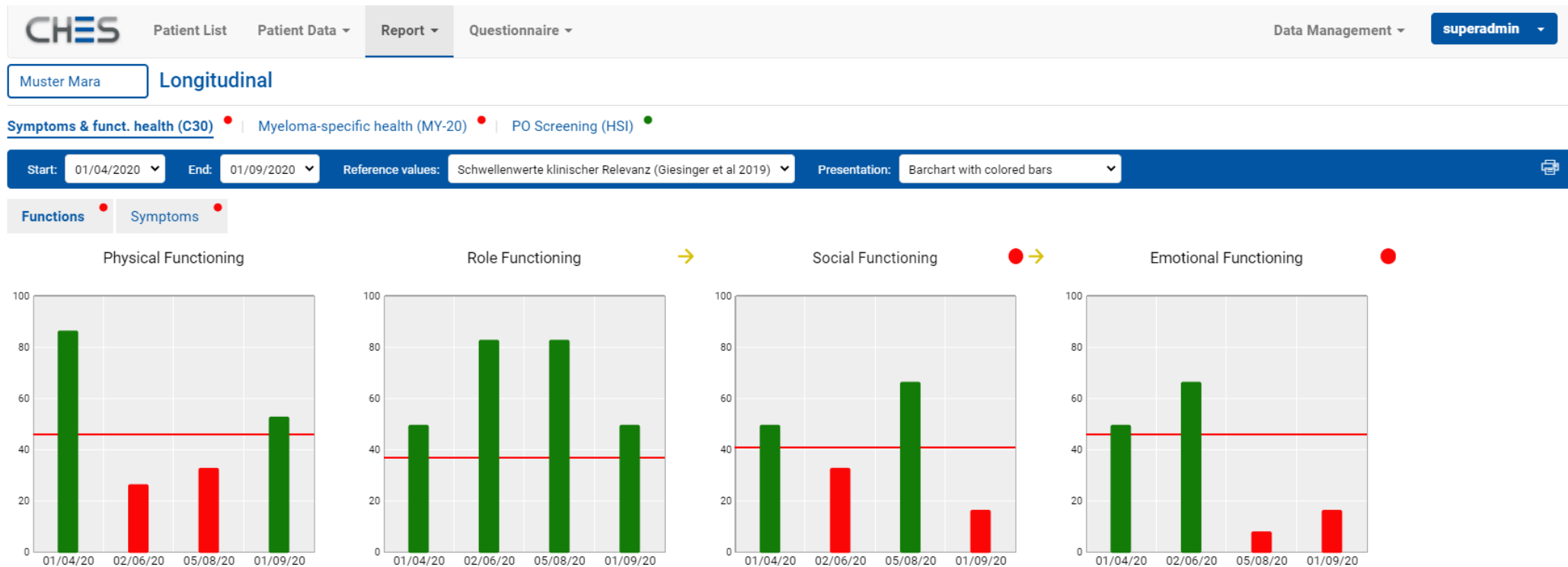
(*) Snyder et al. 2019 Qual Life Res

“For issues specific to individual patient data, the Consensus Panel recommended showing line graphs of scores over time and **including some indication of possibly concerning results in absolute terms** (where evidence exists to support the concerning PRO score range).” (*)

This recommendation fits our threshold methodology and derivation.

It allows these graphs to be used for patient monitoring and management.

Thresholds and graphical result presentation



Screenshot from routine PRO monitoring at the haematological outpatient unit at the Medical University of Innsbruck (Austria)

Implications

- Thresholds for clinical importance support research projects evaluating the clinical benefits of integrating PRO instruments into daily practice and allow to calculate symptom prevalence from PRO scores
- Different thresholds with different meanings may be needed for different purposes
- Symptom prevalence differs depending on underlying definition of what constitutes a “symptom”
- Further research is needed to investigate (in)consistency of thresholds established with different methods
- Thresholds for clinical importance will facilitate score interpretation and contribute to increased acceptance of PRO measures in clinical research and practice



MEDIZINISCHE UNIVERSITÄT
INNSBRUCK

Many thanks for your attention!

johannes.giesinger@i-med.ac.at

Sensitivity analyses

Sensitivity analysis focus on two issues:

- Invariance of **thresholds** across patient groups
- Invariance of **diagnostic accuracy** across patient groups

→ Binary logistic regression analysis

Binary logistic regression model

Dependent variable:

- Binary criterion (derived from anchor items)

Predictor variables:

- QLQ-C30 / EORTC CAT score
- Grouping variable (sex, country, stage, ...)
 - indicates differences in thresholds
- Two-way interaction: Score*Grouping variable
 - indicates differences in diagnostic accuracy

Example I: Model

Do thresholds and/or diagnostic accuracy for **Fatigue** differ for **women and men**?

Binary logistic regression model:

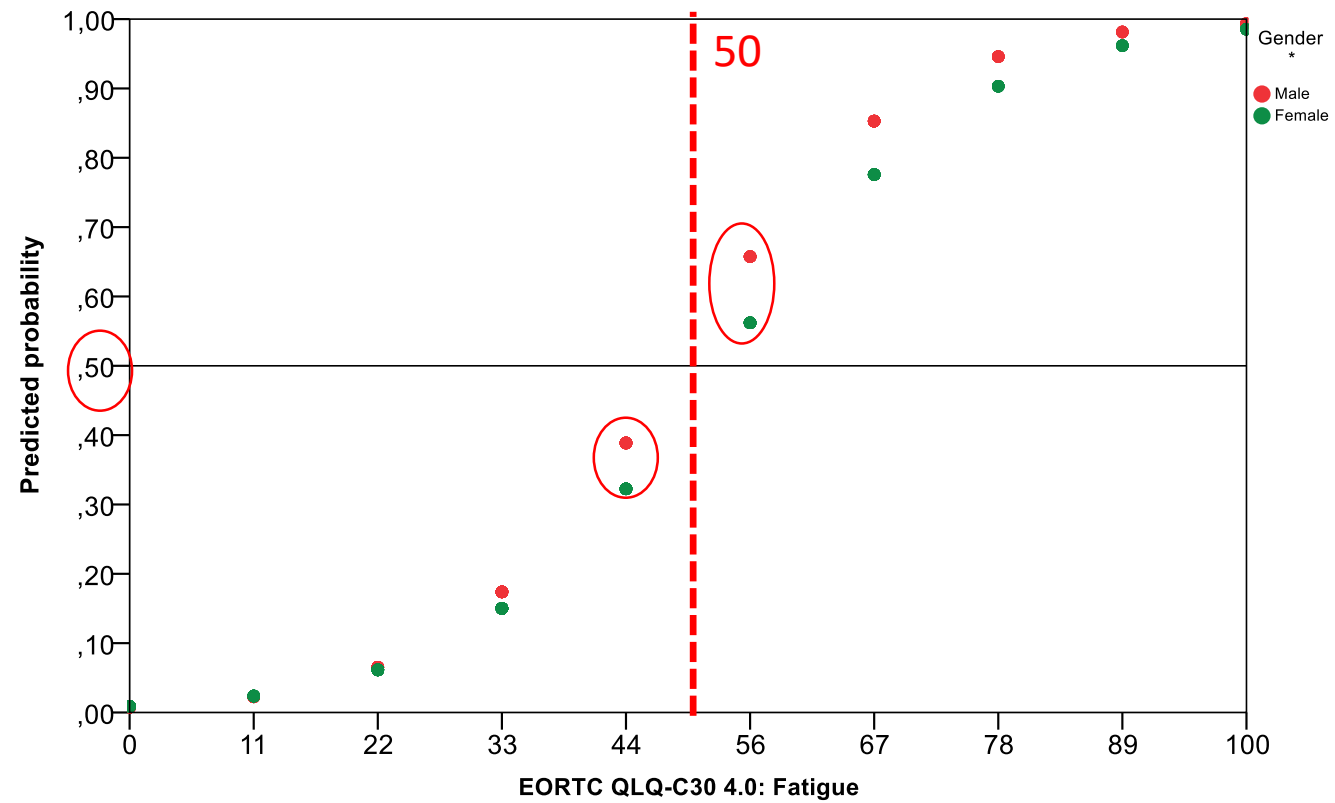
Binary criterion (clinically important yes/no) =

QLQ-C30 FA score + sex + sex*score

Example I: Coefficients

		Variables in the Equation						
		B	S.E.	Wald	df	Sig.	Exp(B)	
Step 1 ^a	EORTC QLQ-C30 4.0: Fatigue	0,089	0,011	68,341	1	0,000	1,093	
	Gender *(1)	-0,164	0,842	0,038	1	0,846	0,849	
	EORTC QLQ-C30 4.0: Fatigue by Gender *(1)	0,010	0,017	0,347	1	0,556	1,010	
	Constant	-4,709	0,539	76,347	1	0,000	0,009	
a. Variable(s) entered on step 1: EORTC QLQ-C30 4.0: Fatigue, Gender *, EORTC QLQ-C30 4.0: Fatigue * Gender * .								

Example I: Plot



Example II: Model

Do thresholds and/or diagnostic accuracy for Physical Functioning differ across countries?

Binary logistic regression model:

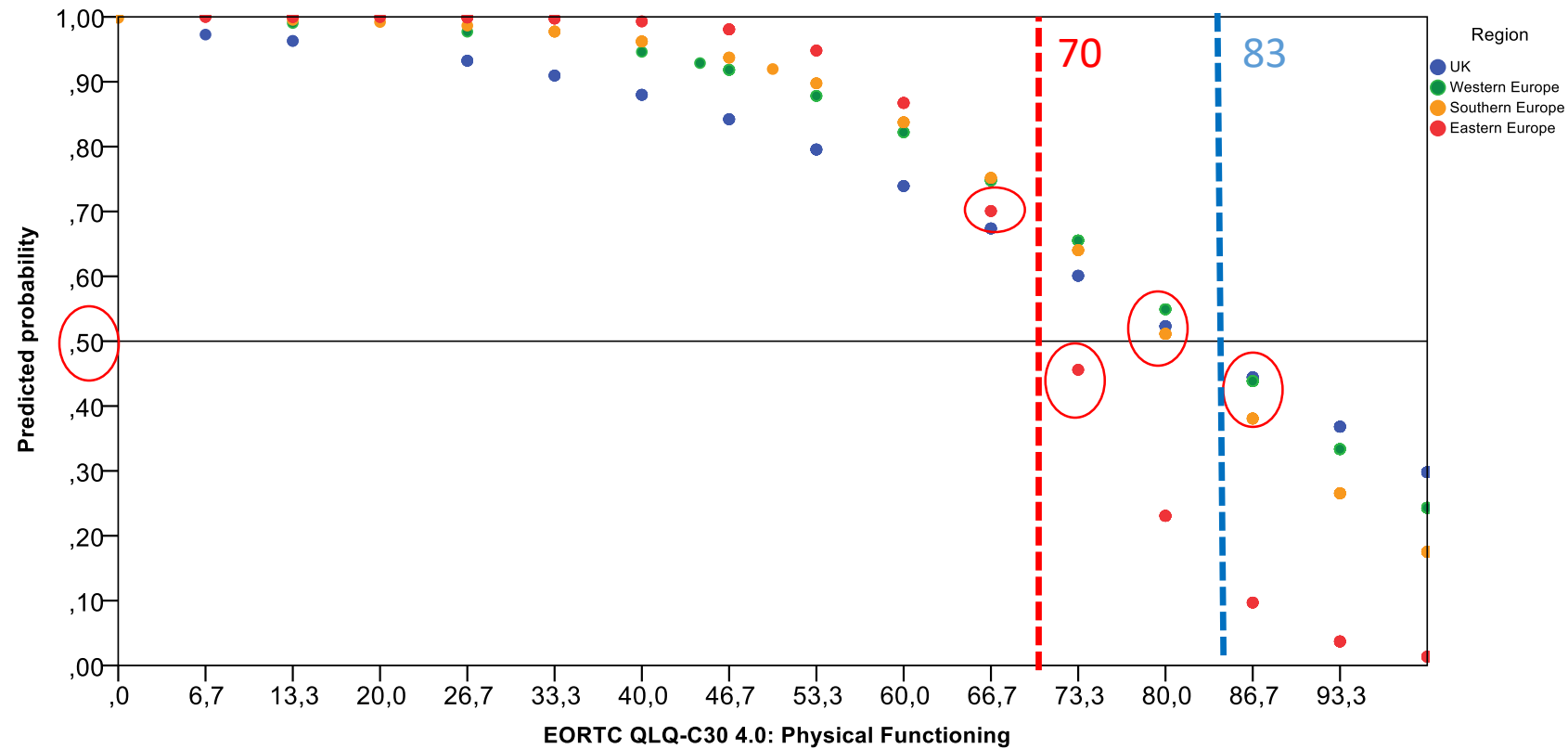
Binary criterion (clinically important yes/no) =

QLQ-C30 PF score + country + score*country

Example II: Coefficients

Logistic Regression Model						
	B	S.E.	Wald	df	Sig.	Exp(B)
Constant	11,123	2,458	20,481	1	0,000	67707,942
EORTC QLQ-C30 4.0: Physical Functioning	-0,154	0,033	22,420	1	0,000	0,857
Region			9,109	3	0,028	
Region(1) Western	-7,233	2,600	7,742	1	0,005	0,001
Region(2) Southern	-5,592	2,737	4,175	1	0,041	0,004
Region(3) Eastern	-4,701	2,799	2,821	1	0,093	0,009
Region *			11,248	3	0,010	
EORTC QLQ-C30 4.0: Physical Functioning						
Region(1) Western by EORTC QLQ-C30 4.0: Physical Functioning	0,107	0,034	9,769	1	0,002	1,113
Region(2) Southern by EORTC QLQ-C30 4.0: Physical Functioning	0,087	0,035	6,065	1	0,014	1,091
Region(3) Eastern by EORTC QLQ-C30 4.0: Physical Functioning	0,074	0,036	4,183	1	0,041	1,077
Region(3) Eastern by EORTC QLQ-C30 4.0: Physical Functioning	0,074	0,036	4,183	1	0,041	1,077
Reference region = UK						

Example II: Plot



Example II: Sens/Spec

Region	Threshold	Sensitivity	Specificity	Youden Index
UK	83	0.70	0.70	0.40
Western Europe	83	0.76	0.69	0.45
Southern Europe	83	0.67	0.80	0.47
Eastern Europe (1)	83	0.61	1.00	0.61
Eastern Europe (2)	70	0.93	0.88	0.81