Aus dem
CharitéCentrum für Innere Medizin und Dermatologie
Medizinische Klinik mit Schwerpunkt Psychosomatik
Direktor: Prof. Dr. med. Matthias Rose

Habilitationsschrift

Advancements in Standardising Patient-Reported Outcomes Measurement

zur Erlangung der Lehrbefähigung
für das Fach Experimentelle Psychosomatische Medizin

vorgelegt dem Fakultätsrat der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von
Dr. Sandra Nolte Graham

Eingereicht: Juni 2022
Dekan: Prof. Dr. med. Axel R. Pries
1. Gutachter: Prof. Dr. Martin Teufel, Tübingen
2. Gutachter: Prof. Dr. Harald Baumeister, Ulm

Table of Contents

Abbreviations

CAT	Computerised adaptive testing
ClinRO	Clinician-reported outcome
COA	Clinical outcome assessment
CTT	Classical test theory
DFG	Deutsche Forschungsgemeinschaft
DIF	Differential item functioning
EMA	European Medicines Agency
EORTC	European Organization for Research and Treatment of Cancer
EU	European Union
FDA	Food and Drug Administration
HRQoL	Health-related quality of life
ICHOM	International Consortium for Health Outcomes Measurement
IRT	Item response theory
ISOQOL	International Society for Quality of Life Research
ISPOR	International Society of Pharmacoeconomics and Outcomes Research
MCID	Minimal clinically important difference
MID	Minimal important difference
NIH	U.S. National Institutes of Health
ObsRO	Observer-reported outcome
PerfO	Performance outcome
PFDD	Patient-Focused Drug Development
PRO	Patient-reported outcome
PROMIS	Patient-Reported Outcomes Measurement Information System
QLG	Quality of Life Group
QoL	Quality of life
RMT	Rasch measurement theory
SD	Standard deviation
SIG	Special Interest Group

1 Introduction

1.1 Patient-Reported Outcomes

The diagnosis of disease as well as evaluations of medical interventions have traditionally relied on standardised objective parameters, such as blood tests or physical examinations. However, particularly in areas where the subjective experience of the patient plays a vital role in understanding the characteristics and course of disease, the integration of the patient perspective provides a more holistic picture of a patient's health status. This is also true for areas where conditions are chronic, and treatments are often aimed at improving or stabilising, rather than curing, the disease. In this case, patients' self-reported symptom experience, their self-reported functioning, or other aspects of the health-related quality of life (HRQoL) may even be the very treatment goal. Hence, patient-reported outcome (PRO) data can deliver crucial information for diagnostic purposes as well as regarding benefits and potential harms of medical interventions. Recognising the potential of the patient's voice (Atherton & Sloan, 2006; Frank et al., 2014), PRO data are increasingly integrated in clinical practice (Ahmed et al., 2012) as well as clinical research (Basch & Leahy, 2019) and they play an increasingly important role in medical and health policy decision making (Porter et al., 2016).

The implementation of routine PRO data collection in clinical practice still frequently relies on local champions with the driving forces being clinicians with support from institutional leadership who recognise the benefits of PROs (Basch, 2017). Such benefits include but are not limited to monitoring the disease course and outcomes assessment. Research has also shown that PRO data improve clinician satisfaction, patient-clinician relationships as well as patient-clinician communication (Basch, 2017; Rotenstein et al., 2017). Moreover, there is strong evidence for the predictive value of PRO data. For example, as part of the Health Survey for England it was found that elevated levels of psychological distress were associated with an increased mortality risk in originally healthy survey participants (Russ et al., 2012). In oncology, PRO data have been found to be predictive of cancer recurrence and cancer survival (De Brabander & Gerits, 1999; Degner & Sloan, 1995; Gotay et al., 2008; Groenvold et al., 2007; Quinten et al., 2014; Sloan et al., 2001). However, despite notable advancements in the integration of PRO data in clinical practice – also stimulated by initiatives such as the International Consortium for Health Outcomes Measurement

(ICHOM) (Porter et al., 2016) – the routine collection of PRO data in clinical practice is still lagging behind advancements in clinical research.

As opposed to clinical practice, the inclusion of PRO data is much more established in clinical research, specifically in clinical trials. Starting with the reflection paper of the European Medicines Agency (EMA) in 2005 (European Medicines Agency, 2005) and the draft PRO Guidance for Industry in 2006 (FDA, 2006) as well as the final PRO Guidance for Industry published in 2009 (FDA, 2009) by the U.S.-American Food and Drug Administration (FDA), both institutions laid the foundations for recognising PROs as crucial endpoints in clinical trials. Over the last 15 years, the FDA in particular has been instrumental in pushing the incorporation of both the patient and caregiver voices in medical product development and it is currently preparing four new guidance documents as part of its Patient-Focused Drug Development (PFDD) Program. These documents aim to provide a systematic approach on how to gather clinical outcome assessment (COA)[1] data and provide guidance on their use. Whilst Guidance 1 that covers sampling methods (FDA, 2020) and Guidance 2 on methods of how to gather COA data from individuals (FDA, 2022) have already been released recently, Guidance 3 on selection/development of fit-for-purpose COA measures and Guidance 4 on incorporating patient experience data in clinical trials are currently under development.

With the growing focus on patient centeredness (Frank et al., 2014) and PRO data becoming an important source for medical and health policy decision making, the requirements concerning the quality of PRO data are increasing as well. And rightly so, as health outcomes researchers and related disciplines working in academia, the pharmaceutical industry and the not-for-profit sector indeed have a growing ethical responsibility to use only those measures that generate robust data and allow for valid inferences from PRO scores (Hawkins et al., 2021; Hawkins et al., 2018; Weinfurt, 2021, 2022). In addition to PRO measures meeting minimum quality criteria (Aaronson et al., 2002; FDA, 2009), it is also crucial that these data are interpretable and can be compared, for example, between patients with similar or different diseases, with general population normative data, across interventions, between cultures as well as across countries. Therefore, standardising PRO measure development and validation processes is vital to ensure the use of PRO data, i.e.,

[1] COAs subsume the following four types of outcome data: patient-reported outcome, clinician-reported outcome (ClinRO), observer-reported outcome (ObsRO) and performance outcome (PerfO) data (FDA, 2020).

the development of an interpretation/use argument (Kane, 1992), is founded on robust validity evidence (Hawkins et al., 2021; Hawkins et al., 2018; Weinfurt, 2021, 2022). At a practical level, it is further essential to provide the users of PROs with guidance on how to interpret the data they obtain.

1.2 Classical and Modern Test Theory Methods

The theory underlying the development and validation of PRO measures borrows from the area of psychometrics that was originally used to develop and evaluate psychological tests (Nunnally & Bernstein, 1994). Psychometric theory is also often referred to as test theory that is short for *"theory of psychological tests and measurements"* (McDonald, 1999). As test theory is concerned with the measurement properties of a scale, it covers the quantitative aspects of the quality of a PRO measure. Test theory is relevant during the PRO measure development/selection process as well as during the measure validation process. In order to assess a PRO measure's psychometric properties, there are two schools of thought that can be drawn upon: 1) classical test theory (CTT) (DeVellis, 2006) and 2) modern test theory, including item response theory (IRT) (Embretson & Reise, 2000) and Rasch measurement theory (RMT) (Andrich, 2011).

The fundamental basis of both CTT and modern test theory approaches is the measurement of latent variables, i.e., unobservable constructs, such as physical (e.g., physical function, pain), mental (e.g., depression, anxiety) or social health (e.g., social support) (Cella et al., 2010). Since latent variables cannot be measured directly, sets of observed variables (items) need to be developed as indicators of the concept of interest. To achieve content validity of a PRO measure, qualitative work involving patients is critical to ensure all aspects of the concept of interest are covered from the patient perspective. The International Society of Pharmacoeconomics and Outcomes Research (ISPOR) Good Research Practices Task Force has published detailed steps of how to develop a PRO measure and is a key reference detailing the qualitative steps to establish a measure's content validity. The steps include definition of the context of use, development of the research protocol, conduct of patient interviews and focus groups, analysis of the qualitative data, and documentation of the concept development, applied methodology and the results (Patrick et al., 2011a, 2011b).

Parallel to the qualitative work, quantitative data in these early stages of content validation need to be gathered as well. For example, these data provide crucial information around

how well the items cover the concept of interest (content coverage), how well the chosen response categories work, including information on the distance between the response categories, the distribution of item and total scores as well as floor or ceiling effects. While these initial quantitative evaluations are exploratory in nature, they lay the groundwork for subsequent confirmatory quantitative approaches as part of the PRO measure validation process (Cappelleri et al., 2014). And it is in the context of these quantitative steps as part of the PRO measure development/selection and validation process that PRO researchers need to decide whether to employ CTT or modern test theory approaches or a combination of the two methods.

Classical Test Theory

Classical test theory is the traditional approach to evaluate the psychometric properties of a PRO measure. It is based on the assumption that each observed score – as obtained from a patient responding to an item of a PRO measure – consists of the hypothetical true score of the unobservable (latent) variable plus random error. Random error means that in CTT it is assumed that the error is unique to a specific item and, therefore, the error is independent of the error on another item. Further, it is assumed that it is just as likely that the error will increase or decrease the observed score with the consequence that errors on the whole cancel each other out, so that the effect of all errors across all respondents on the item's mean score is effectively zero. While the effect on an item's mean score is negligible, error does increase an item's variability (Cappelleri et al., 2014; DeVellis, 2006).

Various types of assessments can be subsumed under CTT, i.e., construct validity, including dimensionality, convergence, discrimination and known groups validity; criterion validity, including concurrent and predictive validity; and reliability, including internal consistency reliability and test-retest reliability (Aaronson et al., 2002; Mokkink et al., 2010). Despite newer test theory approaches, classical test theory has remained popular up to this date. The main reasons are that most scales have been developed based on CTT principles; hence, the underlying measurement model fits most PRO measures reasonably well. Also, items need not be perfect as this potential shortcoming can be offset by other items of the scale. However, because of the fact that adding items to the scale can improve measurement properties, PRO measures in the CTT framework tend to be long. Second, scale scores are typically calculated by using the sum or an average of the individual items. Hence, each item

regardless of its difficulty is given equal weight. Finally, observed score differences around the centre of the score range may have a different meaning compared to the same score difference at the extreme ends of the score range (DeVellis, 2006). Modern test theory approaches are able to offset some of these shortcomings described for CTT methods.

Modern Test Theory

For the purpose of providing the theoretical background on the work presented herein, the main focus of this section is IRT. As opposed to CTT, modern test theory methods – or more specifically IRT – are based on probability theory. Hence, the principles underlying IRT are fundamentally different to those underlying CTT. The basic assumption of IRT is that a respondent's trait level depends on her response to an item and on the item properties. As a result, IRT models are mathematical equations that describe the relationship between a respondent's underlying trait level on the latent variable and the probability of her response to an item that measures the latent variable. IRT models are based on the following two assumptions: 1) unidimensionality, i.e., it is assumed that items contained in an item bank unambiguously measure the target (latent) construct; and 2) local independence, i.e., it is assumed that a person's response to an item is independent of her response to another item (Embretson & Reise, 2000).

The simplest IRT model is a one-parameter logistic model, with the parameter of the model representing item difficulty (severity). For a binary item, i.e., an item with two response categories, the item difficulty parameter indicates the level of the attribute, for example, level of stress or level of anxiety, at which a person has a 50:50 chance of endorsing the item. If the difficulty of the item is higher, it is more difficult for a respondent to endorse the item; hence, a higher level of the attribute is required (Cappelleri et al., 2014). Figure 1 shows an example of physical function with the individual items forming a hypothetical physical function item bank. Following an item calibration process, the physical function items, or questions (Q), of this item bank can be allocated along the continuum of 'physical function', i.e., these items are ordered according to each item's difficulty. As illustrated in Figure 1, the 'easier' items are located further to the left, while 'harder' items are located further to the right of the continuum, with the left anchor being 'low physical function' and the right anchor being 'high physical function'.

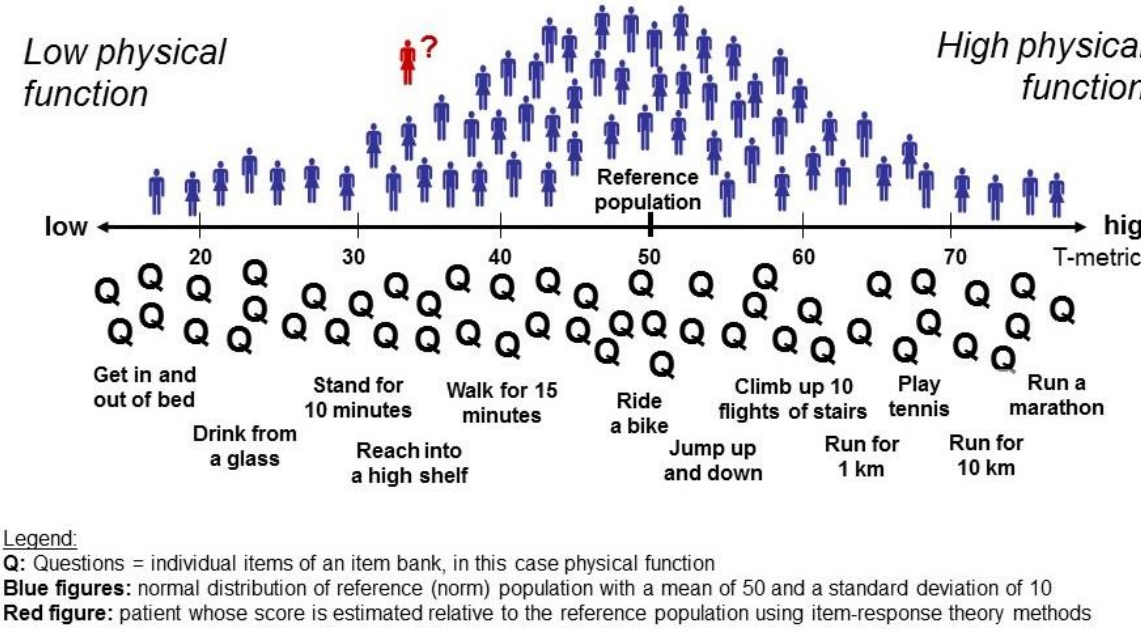

Figure 1. Example of an item bank for physical function

Apart from the one-parameter logistic model, there are two-parameter logistic models that contain a difficulty and an item discrimination parameter. The discrimination – or slope – of an item thereby indicates how well the item discriminates between respondents who have different trait levels or within a respondent if his trait level changes over time. If the slope is steep, a change in trait level has a much bigger impact on the probabilities compared with an item that discriminates less well (Embretson & Reise, 2000), i.e., the steeper the slope, the better the discrimination between persons or within a person over time especially in cases where the score difference is fairly small. Two-parameter logistic models are common in PRO research and are particularly relevant when applying computerised adaptive testing (CAT) (Kocalevent et al., 2009) that is introduced further below. For completeness, there are also three-parameter logistic models that contain the parameters difficulty, discrimination and an additional guessing parameter (Embretson & Reise, 2000); however, these models are uncommon in PRO research. Within the different IRT models, there are a range of models that further vary regarding which item response format they can handle. A good overview of the different types of IRT models as they apply to PRO research can be found in Cappelleri et al. (2014).

A unique feature of IRT models is that – regardless of which items are administered to an individual – the produced scores of the latent traits (theta) can be compared between persons, disease groups, countries, etc. In other words, if two respondents fill out different item subsets of an item bank, their theta scores are still comparable. This assumption forms the basis of CAT (Cella et al., 2010). The idea behind CAT is that it avoids the administration of irrelevant items by tailoring the PRO measure to the individual respondent, which is achieved by making use of the information on individual item parameters (Cella et al., 2007; Gibbons et al., 2008; Rose et al., 2012). Through the individual tailoring of the items, CAT is a more efficient method of PRO assessment, it reduces floor and ceiling effects, and it also reduces respondent burden without compromising the scale's reliability (Cella et al., 2010; Embretson & Reise, 2000).

For a meaningful and sensible interpretation of individual theta scores obtained as part of the IRT framework, a respondent's scores are typically transformed to a T-metric, with a score of 50 defined as the mean of a reference (norm) population and a standard deviation (SD) of 10. Using this approach, scores of individuals can then be interpreted in terms of number of SDs above or below the mean of that reference population. Obtaining data from a reference population is the last step in finalising a PRO measure that has been developed based on modern test theory methods, i.e., before an IRT-based PRO measure is fully functional and can be released to the public, the collection of norm data is crucial. Only if norm data are collected, theta scores can be interpreted meaningfully and sensibly. As opposed to CTT methods, where the comparison of a respondent's score with a reference population is done by looking at the difference between scores, norm data are an integral part of score calculation in IRT, as the estimation of a respondent's score is done in relation to the mean of the reference population.

1.3 International Initiatives Aimed at Standardising PRO Assessment

In light of the growing focus on patient centeredness in medical research (Frank et al., 2014) and PRO data being used as part of medical and health policy decision making (Porter et al., 2016), it is not surprising that a range of efforts around the world have been initiated over the last several years with the aim to standardise PRO assessment, including PRO measure development and validation processes. The various efforts range from initiatives aimed at providing standards for PRO measure development and content validation (Patrick et al.,

2011a, 2011b), minimum standards for the design and selection of PROs (Reeve et al., 2013), standards for the translation and cultural adaptation of PRO measures (Wild et al., 2005), tools for the standardised assessment of PROs (Valderas et al., 2008), guidelines on how to include PROs in clinical trial protocols (Calvert et al., 2018), reporting standards for PROs in protocols and publications (Basch & Leahy, 2019), definitions of key attributes of PRO measures (Aaronson et al., 2002; Mokkink et al., 2010), guides on how to report PRO data in randomised trials (Calvert et al., 2013) to standards for the statistical analysis of PRO data (Coens et al., 2020). Other efforts have been focusing on developing and rigorously evaluating specific PRO instruments and making these available across countries (Aaronson et al., 1992; Aaronson et al., 1993; Cella et al., 2010), whilst initiatives such as ICHOM were established with the sole purpose of defining core outcome sets for a range of indications (Porter et al., 2016).

Two of the initiatives referred to above, as they pertain to the work presented herein, are introduced in more detail below.

One of the most comprehensive efforts to date to standardise the assessment of PRO data is the Patient-Reported Outcomes Measurement Information System (PROMIS®) initiative. It was established in the U.S. in 2004. Since the project's inception, PROMIS collaborators have received public funding from the U.S. National Institutes of Health (NIH) in excess of US$100 million. The purpose of PROMIS is the standardisation of PRO assessment through the development and validation of PRO measures for three major health domains, i.e., mental, physical and social constructs that are aimed to be used across diseases (Cella et al., 2010; Fries et al., 2005). A unique feature of PROMIS is the application of modern test theory methods. Whilst CTT-based instruments are static PRO measures with a fixed number and order of items, the focus of PROMIS is to develop IRT-based item banks that can be used as part of CAT (Fries et al., 2005). In addition to the item banking approach; however, PROMIS has also tested and released various fixed-length short forms (Cella et al., 2019; Segawa et al., 2020).

Another, yet disease-specific, large effort in the standardisation of PRO assessment is the research program of the European Organization for Research and Treatment of Cancer (EORTC) Quality of Life Group (QLG). In 1986, the EORTC initiated the development of a modular approach of assessing the quality of life (QoL) of cancer patients (Aaronson et al., 1993). Starting with the development of a static QoL cancer core questionnaire – the EORTC

QLQ-C30 – the EORTC QLG has since released a large number of modules for several cancer sites that are to be administered alongside the EORTC QLQ-C30 (Sprangers et al., 1993). In 2018, the EORTC QLG also released the EORTC CAT Core, i.e., a CAT version of the EORTC QLQ-C30 using the IRT framework (Petersen et al., 2018; Petersen et al., 2020). To support the interpretation of QoL data obtained from cancer patients, the EORTC QLG published an EORTC QLQ-C30 QoL Reference Values manual in 2008 containing QoL data from various cancer patients stratified by age, gender, cancer stage and cancer site (Scott et al., 2008). The data provided in this manual allows for a comparison of scores obtained from cancer patients with scores from a comparable cancer cohort and therefore greatly enhances the interpretation of EORTC QLQ-C30 data. Apart from reference values, PRO data may also be compared with general population norm data. EORTC QLQ-C30 general population norm data have been collected in various countries over the last decades (Derogar et al., 2012; Hinz et al., 2014; Juul et al., 2014; Mols et al., 2018). However, until rather recently, these norm data were of limited value for application in multi-national clinical cancer trials as no common sampling methodology had been applied across the different studies.

1.4 Research Questions

The overarching theme of the research presented herein relates to different approaches aimed at standardising the measurement and interpretation of PRO data. The research projects were undertaken in the context of the PROMIS and EORTC measurement systems and received funding from Charité – Universitätsmedizin Berlin (Rahel-Hirsch Stipendium), the German Research Foundation (Deutsche Forschungsgemeinschaft [DFG], Sachbeihilfe; Förderzeichen: NO 1138/1-1) and the European Organization for Research and Treatment of Cancer Quality of Life Group (grant number: 001 2015).

The specific aims were to:

1. translate, culturally adapt and undertake an initial psychometric evaluation of the PROMIS Physical Function item bank version 1.2 in German,
2. provide an example of a comprehensive psychometric evaluation of the PROMIS Emotional Distress - Depression item bank version 1.0,
3. generate the European norm for the EORTC CAT Core based on general population norm data obtained from 11 European countries,

4. generate European norm data as well as individual country norm data for a total of 15 countries for the EORTC QLQ-C30, and
5. generate updated German general population norm data for the EORTC QLQ-C30.

2 Original Articles

2.1 Translation, Cultural Adaptation and Psychometric Evaluation of PRO Instruments

PRO instruments are typically developed in one target language as opposed to developing various language versions simultaneously. To be able to use the PRO measure in a language other than the language it was created in – both for national but also for international comparative studies – the measure needs to be translated and culturally adapted.

The study presented here was aimed at the translation and culturally adaptation of the PROMIS v1.2 Physical Function item bank into German and it also carried out an initial psychometric evaluation of the German language version of the item bank applying a combination of CTT and modern test theory methods. Guided by internationally accepted translation principles (Wild et al., 2005), the item bank was first translated and culturally adapted. The study found the German translation to be conceptually equivalent to the original English language version and to be culturally acceptable in the German speaking context. The initial psychometric evaluation showed good psychometric properties with the exception of some of the items of the subdomain 'upper extremity' that did not fully satisfy requirements for a strictly unidimensional construct (Liegl et al., 2018).

The following abstract has been taken from the original peer-reviewed article:

Liegl G, Rose M, Correia H, Fischer HF, Kanlidere S, Mierke A, Obbarius A, Nolte S. An initial psychometric evaluation of the German PROMIS v1.2 Physical Function item bank in patients with a wide range of health conditions. *Clinical Rehabilitation*. 2018;32(1):84–93. DOI: https://doi.org/10.1177/0269215517714297

"OBJECTIVES: To translate the PROMIS Physical Function (PF) item bank version 1.2 into German and to investigate psychometric properties of resulting full bank and seven derived short forms. DESIGN: Cross-sectional psychometric study. SETTING: Inpatient and outpatient clinics of the Department of Psychosomatic Medicine at Charité - Universitätsmedizin Berlin, Germany. SUBJECTS: A total of 10 adult patients with various chronic diseases participated in cognitive debriefing interviews. The final item bank was administered to n = 266 adult patients with a broad range of medical conditions. INTERVENTIONS: Patient-reported outcome assessment as part of routine care. MAIN MEASURES: PROMIS v1.2 PF bank; MOS SF-36 PF scale (PF-10). RESULTS:

Cross-cultural adaptation of the item bank followed established guidelines. For the final German translation, the corrected item-total correlations ranged from 0.44 to 0.84. Cronbach's alpha was high for each PROMIS PF short form (alpha = 0.88-0.96). The full PROMIS PF bank and most short forms correlated highly with the SF-36 PF-10 (r = 0.85-0.90), with the exception of PROMIS Upper Extremity (r = 0.64). PROMIS Upper Extremity showed ceiling effects and lower agreement with the full bank than other short forms. Unidimensionality was supported for all PROMIS PF measures using traditional factor analysis and nonparametric item response theory. CONCLUSION: The German PROMIS PF bank was found to be conceptually equivalent to the English version and fulfilled the psychometric requirements for use of short forms in clinical practice. Future studies should pay particular attention to samples with upper extremity functional limitations to further investigate the dimensional structure of PF as conceptualized according to PROMIS."

2.2 Efforts to Standardise Across Test Theoretical Approaches

The second study was part of a larger effort that had been initiated by members of the Psychometrics Special Interest Group (SIG) of the International Society for Quality of Life Research (ISOQOL). The aim of the effort was to undertake a head-to-head comparison of three different test theoretical approaches, i.e., CTT (Nolte, Coon, et al., 2019), IRT (Stover et al., 2019) and RMT (Cleanthous et al., 2019) and explore whether they would lead to different conclusions about the psychometric performance of a PRO measure. To illustrate the comparison and explore whether the application of different test theoretical methods altered conclusions about a PRO measure's psychometric quality, the PROMIS Depression item bank was used.

As summarised by Björner (Bjorner, 2019), the different approaches showed reasonable agreement between the three approaches regarding the overall conclusions about the psychometric performance of the PROMIS Depression item bank. Largest differences were seen between the basic statistical model, with CTT showing largest differences compared with the results of IRT and RMT that in turn showed reasonable overlap. In contrast, item discrimination was more similar between CTT and IRT, whilst the RMT model suggested exclusion of some of the items from the item bank. Both the analysis of the item thresholds (difficulty parameter) and differential item functioning (DIF) showed comparable results between IRT and RMT; the CTT study did not explore either of the two.

In conclusion, the CTT paper presented below provided an important contribution to this collaborative effort and suggested that – regardless of the test theoretical approach taken – the overall conclusion about the psychometric quality of the PROMIS Depression item bank was sufficiently similar between CTT, IRT and RMT. The CTT paper further provided a basic introduction to the various psychometric tests undertaken within the CTT framework.

The following abstract has been taken from the original peer-reviewed article:

Nolte S, Coon C, Hudgens S, Verdam MGE. Psychometric evaluation of the PROMIS® Depression Item Bank: an illustration of classical test theory methods. *Journal of Patient-Reported Outcomes*. 2019;3(1):46. DOI: https://doi.org/10.1186/s41687-019-0127-0

> *"BACKGROUND: Psychometric theory offers a range of tests that can be used as supportive evidence of both validity and reliability of instruments aimed at measuring patient-reported outcomes (PRO). The aim of this paper is to illustrate psychometric*

tests within the Classical Test Theory (CTT) framework, comprising indices that are frequently applied to assess item- and scale-level psychometric properties of PRO instruments. METHODS: Using data on the PROMIS Depression Item Bank, typical CTT indices for the assessment of psychometric properties are illustrated, including content validity, item-level data exploration, reliability, and construct validity, particularly confirmatory factor analysis, to test the unidimensionality assumption underlying the item bank. Analyses are carried out on an original item set of 51 depression items, the final (official) PROMIS Depression Item Bank consisting of 28 items, and an 8-item short form. RESULTS: The analyses reported provide an informative illustration on how item- and scale-level reliability and validity statistics can be used to assess the psychometric quality of a PRO instrument. The results illustrate how the reported statistics can be used for item selection from an item pool (here: 51 items). Both the (final) 28-item bank and the 8-item short form show good psychometric properties supporting the high quality of individual items and the unidimensionality assumption of the item bank. CONCLUSIONS: It is our hope that our illustration of CTT methods, in conjunction with two companion papers illustrating modern test theory methods, will help researchers to confidently apply a range of statistical tests to evaluate item- and scale-level psychometric performance of PRO instruments."

Journal of Patient-Reported Outcomes

RESEARCH Open Access

Psychometric evaluation of the PROMIS® Depression Item Bank: an illustration of classical test theory methods

Sandra Nolte[1,2], Cheryl Coon[3], Stacie Hudgens[4*] and Mathilde G. E. Verdam[5,6]

Abstract

Background: Psychometric theory offers a range of tests that can be used as supportive evidence of both validity and reliability of instruments aimed at measuring patient-reported outcomes (PRO). The aim of this paper is to illustrate psychometric tests within the Classical Test Theory (CTT) framework, comprising indices that are frequently applied to assess item- and scale-level psychometric properties of PRO instruments.

Methods: Using data on the PROMIS Depression Item Bank, typical CTT indices for the assessment of psychometric properties are illustrated, including content validity, item-level data exploration, reliability, and construct validity, particularly confirmatory factor analysis, to test the unidimensionality assumption underlying the item bank. Analyses are carried out on an original item set of 51 depression items, the final (official) PROMIS Depression Item Bank consisting of 28 items, and an 8-item short form.

Results: The analyses reported provide an informative illustration on how item- and scale-level reliability and validity statistics can be used to assess the psychometric quality of a PRO instrument. The results illustrate how the reported statistics can be used for item selection from an item pool (*here*: 51 items). Both the (final) 28-item bank and the 8-item short form show good psychometric properties supporting the high quality of individual items and the unidimensionality assumption of the item bank.

Conclusions: It is our hope that our illustration of CTT methods, in conjunction with two companion papers illustrating modern test theory methods, will help researchers to confidently apply a range of statistical tests to evaluate item- and scale-level psychometric performance of PRO instruments.

Keywords: Classical test theory, Patient-reported outcomes, Validity, Reliability, Factor analysis, Structural equation modeling

Background

Test theory, also referred to as psychometric theory, is concerned with the theory of measurement of psychological constructs [1]. Although initial developments of test theory date back more than a century [2], psychometric theory is more topical than ever, in particular in the field of medicine. Over the past decade, the inclusion of the patient perspective in clinical care and research (e.g., through measurement of self-reported outcomes such as symptom burden, emotional, physical, and social functioning) has developed to be a necessary rather than merely desired aspect in the evaluation of treatment effectiveness, with regulatory agencies worldwide recommending the inclusion of patient-reported outcomes (PROs) in clinical trials [3–5]. The growing importance of patient-centeredness, not only in the delivery of healthcare but also in healthcare research, is further noticeable in the increased funding dedicated to both improvement and standardization of PRO measures. For example, the Patient-Centered Outcomes Research Institute (PCORI) was founded in the United States in 2010 with the aim to fund only those comparative effectiveness research studies that demonstrate engagement with and to be of relevance to patients and caregivers [6].

* Correspondence: stacie.hudgens@clinoutsolutions.com
[4]Clinical Outcomes Solutions, 1820 East River Road, Suite 220, Tucson, AZ 85718, USA
Full list of author information is available at the end of the article

Further, the standardization of PRO assessment has become a major research area; initiatives, such as the Patient-Reported Outcomes Measurement Information System (PROMIS®), have been founded to develop and validate item banks on major health domains that are successively being implemented across the globe [7, 8].

In view of increased use and relevance of PROs, it is crucial that these self-reported outcomes are measured with the utmost precision. For this, psychometric theory is pivotal as it offers a range of tests that can be used as supportive evidence of both validity and reliability of a PRO instrument. In other words, because the psychological phenomenon of interest cannot be observed directly (e.g., depression), it is necessary to assess the extent to which the self-report measure (i.e., the set of items on a questionnaire) can be interpreted as a valid and reliable reflection of the construct that it is intended to measure. As such, psychometric theory plays an important role in the development of PRO instruments and the evaluation of their psychometric quality.

Both traditional and modern test theory methods can be employed to evaluate an instrument's psychometric properties. At the core of both methods is that they are concerned with the measurement of unobservable (latent) constructs through a set of observed variables to get as best an approximation of the latent variable as possible. Traditional test theory, also referred to as Classical Test Theory (CTT), is the older of the two and still the most frequently applied method in health-related quality of life research; its use is also suggested by the U.S.-American Food and Drug Administration [4]. Generally, CTT includes indices that describe a PRO instrument's validity (content/face, construct [structural, convergent, discriminant, and known groups], criterion [concurrent and predictive]) and its reliability (internal consistency and test-retest reliability) [9, 10].

The aim of this paper is to provide an illustration of a range of analyses within the CTT framework. We discuss both the (practical) advantages and disadvantages of the analyses and their interpretation. This educational paper is part of a series of papers initiated by the Psychometrics Special Interest Group (SIG) of the International Society for Quality of Life Research (ISOQOL) aimed at introducing different psychometric techniques to analyze item properties of a PRO instrument, i.e., CTT as presented in this paper, item response theory (IRT) [Stover et al, copublished in this issue], and Rasch measurement theory (RMT) [Cleanthous et al, copublished in this issue] methods. To outline the methods used to perform psychometric tests applying a CTT-based approach, the PROMIS® Emotional Distress - Depression Item Bank version 1.0 was selected because of its availability and extensive use since its development in 2011 [11]. Although other PRO instruments are available to assess depression (e.g., Center for Epidemiological Studies – Depression [CES-D] [12], Patient Health Questionnaire [PHQ-9] [13], Beck Depression Inventory II [BDI-II)] [14]), the PROMIS Depression Item Bank has been shown to provide more information than conventional measures for which these short-form measures are comparable [11]. The objective of the present paper is to use data on the PROMIS Depression Item Bank to demonstrate how CTT methods may be employed to evaluate the psychometric properties of a set of PRO items.

Methods

PROMIS emotional distress - Depression Item Bank version 1.0

The PROMIS Depression Item Bank was developed following a comprehensive literature search and qualitative methods, which resulted in an initial pool of 518 items. Using psychometric analyses, these were subsequently reduced to a preliminary pool of 56 items for calibration testing. After thorough quantitative analyses, the final PROMIS Depression Item Bank contains 28 items [11]. The items included in the final bank specifically focus on negative mood, decreases in positive emotions, cognitive deficits, negative self-image, and negative social cognition [7]. The items are scored on a 5-point verbal response scale (i.e., ordered categorical item responses) where respondents are asked to rate the experienced frequency of symptoms (*never*, *rarely*, *sometimes*, *often*, *always*).

For the purpose of this series of papers, a subset of the PROMIS calibration samples was made available by the PROMIS Health Organization. This dataset comprised 51 of the 56 preliminary PROMIS depression items, and was seen as a valuable resource in the public domain by aforementioned ISOQOL Psychometrics SIG to compare item performance results from CTT, IRT, and RMT methods. The full sample data is also publicly available (see https://doi.org/10.7910/DVN/0NGAKG).

The PROMIS Calibration Studies sample included 21,133 respondents, with $n = 1532$ recruited from primary research sites associated with PROMIS network sites, while the vast majority ($n = 19{,}601$) was recruited from an Internet polling company; further details about the sampling are available in the introductory paper to this special issue [ref]. For the purpose of illustrating different methods to assess the psychometric properties, we only used data from respondents from the general population that were administered the full item bank ($n = 925$), and excluded respondents that were flagged by predetermined speed-of-response criteria ($n = 100$) and who had missing item responses ($n = 72$). This resulted in a total sample of $N = 753$ (see Table 1 for an overview of demographic information). Pilkonis et al. [11] have

Table 1 Demographic Characteristics of the PROMIS Sample ($N = 753$)

	General population sample ($N = 753$)
Age; Mean (SD)	51 (19)
Age group; N (%)	
18–35	204 (27)
36–50	164 (22)
51–65	182 (24)
66–88	198 (26)
Gender; N (%)	
Female	391 (52)
Male	361 (48)
Ethnicity; N (%)	
Caucasian	597 (79)
African-American	73 (10)
Other	83 (11)
Education; N (%)	
Primary	20 (2)
Secondary	149 (20)
Post-secondary	346 (46)
Tertiary	238 (32)
Relationship status; N (%)	
Single	120 (16)
Married or with relationship	485 (64)
Separated or divorced	87 (11)
Widowed	59 (8)

previously described results of psychometric analyses on the same data for the purpose of item selection. The current paper does not have such a substantive aim – we do not wish to add to the analyses reported of Pilkonis et al. [11] – but rather our aim is to use these data for illustrative purposes to introduce CTT methods. As the final item bank contains 28 items, which can further be applied as one of the many PROMIS depression short forms (in this case, 8-item Short Form 8b), subsequent analyses were carried out on three item subsets, i.e. 51, 28, and 8 items, respectively.

Content/face validity

Content validity refers to the extent to which a questionnaire's items reflect the content of the construct to be measured. Establishing content validity is a theoretical and subjective undertaking that is part of the instrument development process; although it is an important psychometric quality it does not require statistical evaluation. It is done by providing a definition of the target construct (e.g., using literature search, focus groups, interviews) and subsequent development of new items and/or selection of items from existing instruments. A related concept is face validity, which refers to the extent to which experts agree on what the instrument *appears* to measure. The main distinction between the two is that content validity refers to the instrument development process, whereas the latter term is usually used in the context of critical review of existing instruments [15].

Data exploration at the item level

Although CTT techniques generally focus on tests at the scale level, it is useful to include item-level exploration in the evaluation of an instrument's measurement properties. This can be done by, for example, inspecting frequency distributions (ordinal item responses) or means and variances (continuous item responses). Generally, variability across response categories and items is desirable as it indicates that respective item's content is relevant to respondents and the response categories are appropriate for determining the continuum of a psychological construct. The distribution of responses also gives insight into potential floor or ceiling effects. While item-level exploration can give important insight into the quality of individual items, strict decision rules regarding whether response variability is adequate is difficult given that frequency distributions/means (variances) are dependent on the construct being measured and the sample used. For example, in a general population sample, response variability on items about severe depression symptoms is expected to be limited as compared to mild depression symptoms, whereas one may expect the reverse in a clinical sample. Additionally, a clinical sample participating in a clinical trial is likely to demonstrate quite different item response distributions at baseline (i.e., when they are symptomatic and in need of treatment) versus post-treatment (i.e., when the treatment has hopefully improved symptoms). Therefore, contextual factors should be considered when interpreting item-level analyses; one may accept different distributions in different samples under different conditions as reasonable. Given the ordinal scaling of the PROMIS depression items, in this paper the frequency distribution for each item was examined to evaluate data completeness, potential floor and/or ceiling effects, and the variability of responses across categories.

Item discrimination refers to the extent to which an item measures the underlying construct of interest, and thus is able to discriminate between respondents. Item discrimination is determined by exploring the correlation of an individual item with the whole item set (item-total correlation) or with all other items of the set (corrected item-total correlation). In this paper, we considered corrected item-total correlations of $r_{itc} < 0.4$ as the cut-off following the developers of the PROMIS

Depression Item Bank [11], but other cut-offs have been suggested (see [16]).

Aforementioned distinction regarding whether items are scored on an ordinal (e.g., Likert) or continuous response scale also influences the choice of further statistical methods used to inspect an instrument's quality. As the former type can only take on a limited number of values, a decision has to be made about whether these can be treated as (approximations of) continuous item responses or as ordered categories. Inspection of skewness/kurtosis statistics (transformed to a *z*-score) and normality tests (e.g., Kolmogorov-Smirnov test, Shapiro-Wilk test) can be used to evaluate the assumption of normal distribution of item scores. However, with larger samples ($n > 200$) these tests can turn out to be significant even with only small deviations from normality. As an alternative, visual inspection of the distribution or substantive considerations may be more appropriate to guide a decision regarding which statistical methods to use [17].

Reliability

Reliability refers to the extent to which the scores on an item set reflect the 'true' score on the construct of interest. Scores that are highly reliable are accurate, reproducible, and consistent reflections of the underlying construct that the item set measures. Different methods exist to evaluate scale reliability where the reliability coefficient reflects the proportion of true variance in the variance of the observed scores, with higher values indicating a more reliable estimate of the true scores.

The most well-known and widely applied reliability coefficient is Cronbach's alpha [18], also referred to as a measure of internal consistency. Values > 0.70 are generally taken to indicate good reliability; however, the appropriateness of this – or any – threshold may vary depending on the purpose of the instrument [19]. Although Cronbach's alpha is most often used as a reliability coefficient, it is not without critique [20–22]. In particular, it is estimated under the assumption that all items are equally good measures of the construct (i.e., they are essentially tau equivalent) and violation of this assumption may lead to an underestimation of the reliability. Moreover, its calculation is influenced by the number of items of the test and the average interrelatedness between the items, which may result in high reliability estimates for longer tests regardless of whether the items measure a homogeneous (i.e., unidimensional) construct or not. Therefore, interpretation of Cronbach's alpha should coincide with a careful consideration of both the instrument's content and number of items in the scale. In order to take into account deviating distributional properties, it may be more appropriate to apply alternatives that have been developed, such as a special correction to the reliability coefficient that has been suggested for the ordinal case [23] and the Kuder-Richardson Formula (K-R20) that is a simplified version of Cronbach's alpha for the dichotomous case [24]. The internal consistency of the PROMIS Depression Item Bank was assessed using the alpha coefficient with a threshold criterion of > 0.70 [16].

Construct validity

Construct validity refers to the extent to which the behavior of the instrument's scores are consistent with what would be expected from the construct of interest. This can be evaluated by looking at internal relationships, relationships to scores of other instruments or differences between relevant groups.

In the following we address construct validity in terms of internal relationships between variables (i.e., dimensionality/structural validity) using factor analysis [10]. Further analyses of construct validity are considered outside the scope of the current article. As an example, in the context of the PROMIS Depression Item Bank one could consider further investigation of construct validity by looking at the correlations with different measures of depression, or investigate differences in depression scores between a clinical sample and general population sample.

Dimensionality

Assessment of an instrument's dimensionality is also referred to as structural validity. It is used to assess the degree to which the scores of an item set are an adequate reflection of the dimensionality of the construct. Within the CTT framework, structural validity is usually assessed using factor analysis. The factor analytic framework is historically closely connected to the CTT framework, although it can be considered as a more 'modern' set of statistical techniques as it allows for the investigation of item- and scale-characteristics of an instrument using less restrictive assumptions. That is, the flexibility of the factor analytic framework can be used to model the individual item characteristics without imposing equality restrictions (i.e., assuming tau equivalences or parallelism of the items).

Factor analysis is a group-level analysis technique aimed at attributing sets of observed variables to one or more latent variables [25]. While exploratory factor analysis is generally used when the relationship among the variables is unknown and the researcher is seeking dimensionality insight from the analysis, confirmatory factor analysis (CFA) is more appropriate when relationships among the variables are already hypothesized (e.g., via a conceptual framework used to construct the instrument). In CFA, the more variance of an item can be explained by the hypothesized latent variable (factor), the better the item fits to

the construct. This is usually expressed in terms of *factor loadings*, with the squared loading indicating the variance explained. Loadings > 0.50 are deemed the minimum; loadings > 0.70 are desirable [19]. To confirm the hypothesized one-factor structure of the PROMIS Depression Item Bank, in this paper unidimensional CFA was used to assess the degree to which it is appropriate to combine the 51, 28, and 8 items, respectively, in one domain [26].

To evaluate how well the hypothesized model fits the data the most widely used method is maximum likelihood (ML) estimation. It is valid under the assumption that observed scores follow a multivariate normal distribution; however, alternative estimation methods are required when this assumption is not met (e.g., with ordinal data [27, 28]). Options for ordinal data include weighted least squares for large samples and simple models, and robust (or diagonally) weighted least squares (WLSMV/DWLS) for smaller samples and complex models. These estimation methods use an asymptotic covariance weight matrix to adjust for the non-normality of ordinal data [29]. In addition, estimation methods for ordinal data usually require adjustments to the input matrices of variances and covariances. That is, in the ordinal and dichotomous case, polychoric and tetrachoric correlations, respectively, are estimated instead.

To evaluate overall goodness-of-fit, the χ^2 test of exact fit [30] can be used where a significant χ^2 value indicates a significant difference between data and model [26]. As this value is dependent on sample size and number of model parameters included [26, 31], alternative fit indices have been developed. A prominent approximate fit index is the root mean square error of approximation (RMSEA), with RMSEA≤.05 indicating close and RMSEA≤.08 indicating reasonable approximate fit [32]. Finally, incremental fit indices are used to compare the model to an alternative or baseline model [33], with the comparative fit index (CFI) [34] most frequently recommended [35]. CFI values range between 0 and 1 [19, 34]; CFI ≥ 0.95 is indicative of good model fit [36]. The Tucker Lewis Index (TLI) – or non-normed fit index (NNFI) – is conceptually similar to CFI. While not normed between 0 and 1, values close to 1 are considered to indicate good fit [19]. For more detailed overviews of different fit indices and their interpretation, the reader may be referred elsewhere [36–38].

In this paper, CFA was conducted using polychoric correlations and WLSMV with robust standard errors and a mean- and variance-adjusted test statistic (using a scale-shifted approach). Further, above fit indices based on the adjusted chi-square test statistic were considered to interpret goodness of model fit. In the event that the unidimensional model did not provide acceptable model fit, modification indices (i.e., the expected change in model fit if specific model revisions were made) and residual correlations (i.e., the excess relationship between items after accounting for the underlying factor) were inspected to identify reasons for model misfit [39]. Analyses were conducted with the package Lavaan that runs in the freely available R software [40]. Syntaxes of the analyses are available on request.

Results

Content/face validity

As content validity of the PROMIS Depression Item Bank was performed by the original developers, it is only presented here for completeness [41]. It comprised a comprehensive literature search [41] and focus groups with patients to ensure that the instrument reflected the perspectives of the population of interest [42]. Moreover, selection of items was (partly) based on content balancing to retain a representative group of symptoms and complaints in the final bank. Face validity was assessed by asking experts to review the resulting bank, and to define and describe the content that was being measured [43].

Data exploration at the item level

Normality tests showed that all items deviated significantly from normality, with severe right skewness (Table 1). Visual inspection of histograms and frequency distributions confirmed that response options 'never' and 'rarely' were chosen more frequently (e.g., regularly > 60% of respondents) than the other response options. It needs to be taken into account that the sample used consisted of a representative population sample, which may explain the relatively low percentages of endorsement of the more extreme response categories. The finding that these items show low response variability is thus limited for administration in a general population sample, as the behavior of items could be quite different in other, e.g. clinical, samples. Dependent on the intended use of the instrument, these results could serve as basis for item selection by removing items that show the most skewed item response distributions (as was also done by Pilkonis et al. [11]). For example, one could remove item 1 ('I reacted slowly to things that were said or done') as only 3.5% fall within the two highest response categories. Alternatively, item 15 ('I disliked the way my body looked') seems to show almost a uniform response distribution, with similar percentages of respondents in each response category. This pattern of responses deviates from the other response patterns and may indicate that this item measures something else than depression (as measured by the other items). In contrast, an item such as item 32 'I wished I were dead and away from it all', where 84% of respondents chose response option 'never' and only 1% indicated 'always', could be retained based on item content, as it could be deemed a relevant item to cover suicidal thoughts. However, if the

instrument was to be used in a clinical population, then we may want to be more conservative with item reduction, as these items with low endorsement in a general population may be relevant in a clinical population and important for measuring severe conditions.

Based on these distributional results, taken in combination with the nature of the response options (i.e., frequencies), we decided that the PROMIS depression items should be treated as ordinal, although observed variables with five response categories are sometimes treated as continuous.

The corrected item-total correlations were investigated next (three rightmost columns of Table 1) using polychoric correlations to take into account the ordinal nature of the data, as suggested by Gaderman, Guhn and Zumbo [23]. There was one item with correlation smaller than 0.40 (i.e., item 49 'I lost weight without trying'), which could therefore be considered for removal. As all other corrected item-total correlations were rather high, one could consider using more stringent criteria for item selection. For example, items 11, 15 and 53 could be removed based on the criterion that the corrected item-total correlation should be larger than 0.70, and an additional 8 items would be candidates for removal if the criterion were increased to 0.75 (items 1, 3, 18, 20, 24, 34, 37, and 43). These candidate items for removal are consistent with the item selection of Pilkonis et al. [11]). Thus, item-level data exploration can provide valuable information about the performance of items within a scale; however, it should be used in combination with further substantive considerations to retain a sensible item set.

Reliability

The reliability coefficient was calculated based on polychoric correlations, as appropriate for ordinal data [23]. The alpha reliability coefficient was high, with 0.989 for the total item bank (51 items), 0.988 for the final item bank (28 items), and 0.974 for the 8-item short form, with the latter finding indicating that reliable measurement of depression can be attained with a relatively small number of items. To consider item reduction based on the reliability coefficient, one could further inspect the alpha-if-item-deleted statistic (i.e., the expected alpha of the instrument when the specific item is deleted), which identifies items that may not be highly related to the other items and the domain of interest (depression). For example, the deletion of aforementioned item 49 from the 51-item bank would not substantially change alpha (0.990 versus 0.989), suggesting that this item was somewhat different from the other 50 items.

Construct validity - dimensionality

The one-factor solution for the 51-item bank showed acceptable model fit (Table 2). All standardized factor loadings were quite high (mostly > 0.7; Table 3), thus supporting the unidimensionality assumption of the underlying depression factor. Only for item 49 the factor loading was low; hence, again a candidate item for deletion. Using more stringent criteria, e.g. factor loadings of ≥0.7, there are three additional candidate items for deletion. Inspection of modification indices showed that the three most problematic instances of misfit in the model were a result of high residual correlations between items 16 ('I felt like crying') and 34 ('I had crying spells'), items 3 ('I felt that I had no energy') and 18 ('I got tired more easily than usual'), and items 32 ('I wished I were dead and away from it all') and 39 ('I felt I had no reason for living'). These results indicate that, in this general population, the respective relationship between these items was not well explained by the underlying factor. In other words, these items have something else in common besides what is captured by the depression factor. Closer inspection of respective content suggests that each item pair seem to measure similar symptoms. Such finding could be an indication of multidimensionality as items that belong to the same subdomain could be considered to reflect a multidimensional depression construct; alternatively, it could be an indication of item redundancy, i.e., keeping both items of respective item pair in the bank does not add substantial new information so that one item may be removed. This hypothesis could be further explored either by looking at patterns of residual correlations or by fitting multidimensional factor models.

Results of the CFA containing the 28 items of the final PROMIS Depression Item Bank generally supported the unidimensionality assumption as did the 8-item short

Table 2 Model-fit results of confirmatory factor analysis (CFA) of extended item bank (51 items), final item bank (28 items) and short-form 8b (8 items) of the PROMIS Depression Item Bank using WLSMV[a,b]

Model	χ^2 value	df	*p*	RMSEA [90% CI]	CFI	TLI
Unidimensional model with 51 items	5729.6	1224	<.001	0.070 [0.068; 0.072]	0.953	0.951
Unidimensional model with 28 items	1473.27	350	<.001	0.065 [0.062; 0.069]	0.983	0.982
Unidimensional model with 8 items	177.71	20	<.001	0.101 [0.088; 0.115]	0.995	0.992

[a]WLSMV: robust weighted least squares
[b]Sample size for the models with 51, 28 and 8 items, respectively, was N = 753

Table 3 Factor loadings for the unidimensional factor models of the PROMIS Depression Item Bank

51 items	Factor loading	Residual variance	28 items	Factor loading	Residual variance	8 items	Factor loading	Residual variance
EDDEP01	0.733	0.462	EDDEP04	0.928	0.139	EDDEP04	0.924	0.147
EDDEP03	0.763	0.418	EDDEP05	0.9019	0.188	EDDEP05	0.888	0.212
EDDEP04	0.922	0.151	EDDEP06	0.901	0.189	EDDEP06	0.910	0.172
EDDEP05	0.893	0.203	EDDEP07	0.825	0.319	EDDEP17	0.891	0.206
EDDEP06	0.894	0.200	EDDEP09	0.900	0.190	EDDEP22	0.905	0.180
EDDEP07	0.826	0.317	EDDEP14	0.852	0.274	EDDEP29	0.900	0.190
EDDEP08	0.793	0.371	EDDEP17	0.877	0.232	EDDEP36	0.925	0.145
EDDEP09	0.897	0.195	EDDEP19	0.904	0.182	EDDEP41	0.944	0.109
EDDEP11	0.530	0.719	EDDEP21	0.846	0.284			
EDDEP12	0.796	0.366	EDDEP22	0.918	0.158			
EDDEP14	0.843	0.289	EDDEP23	0.823	0.323			
EDDEP15	0.609	0.629	EDDEP26	0.868	0.247			
EDDEP16	0.820	0.328	EDDEP27	0.870	0.243			
EDDEP17	0.872	0.239	EDDEP28	0.823	0.323			
EDDEP18	0.770	0.407	EDDEP29	0.898	0.194			
EDDEP19	0.901	0.188	EDDEP30	0.802	0.357			
EDDEP20	0.773	0.403	EDDEP31	0.863	0.254			
EDDEP21	0.835	0.304	EDDEP35	0.861	0.259			
EDDEP22	0.904	0.183	EDDEP36	0.911	0.169			
EDDEP23	0.816	0.335	EDDEP39	0.879	0.226			
EDDEP24	0.710	0.496	EDDEP41	0.938	0.120			
EDDEP26	0.853	0.272	EDDEP42	0.799	0.362			
EDDEP27	0.861	0.258	EDDEP44	0.827	0.317			
EDDEP28	0.813	0.339	EDDEP45	0.832	0.308			
EDDEP29	0.901	0.189	EDDEP46	0.890	0.308			
EDDEP30	0.826	0.318	EDDEP48	0.880	0.226			
EDDEP31	0.848	0.281	EDDEP50	0.795	0.368			
EDDEP32	0.891	0.206	EDDEP54	0.859	0.261			
EDDEP33	0.817	0.332						
EDDEP34	0.779	0.394						
EDDEP35	0.863	0.255						
EDDEP36	0.900	0.190						
EDDEP37	0.726	0.472						
EDDEP38	0.802	0.356						
EDDEP39	0.904	0.183						
EDDEP40	0.841	0.293						
EDDEP41	0.927	0.140						
EDDEP42	0.792	0.373						
EDDEP43	0.776	0.398						
EDDEP44	0.838	0.298						
EDDEP45	0.835	0.303						
EDDEP46	0.819	0.330						
EDDEP47	0.800	0.360						
EDDEP48	0.871	0.241						

Table 3 Factor loadings for the unidimensional factor models of the PROMIS Depression Item Bank *(Continued)*

51 items	Factor loading	Residual variance	28 items	Factor loading	Residual variance	8 items	Factor loading	Residual variance
EDDEP49	0.300	0.910						
EDDEP50	0.788	0.380						
EDDEP52	0.838	0.298						
EDDEP53	0.614	0.623						
EDDEP54	0.874	0.236						
EDDEP55	0.860	0.260						
EDDEP56	0.846	0.283						

form, with the latter showing somewhat less optimal model fit (Table 2). Further, standardized factor loadings for all items were very high, with associated low proportions of unexplained item variances (Table 3).

Discussion

This paper aimed to demonstrate how CTT methods may be employed to evaluate the psychometric properties of a set of PRO items. It is part of a series of papers initiated by the ISOQOL Psychometrics SIG, which was aimed at comparing CTT with two modern test theory approaches, i.e. IRT and RMT methods, with each using the same dataset of PROMIS depression items.

First, item-level analyses were carried out. As the data made available were collected from a general population, the prevalence of the various depression symptoms was rather low, which resulted in high floor effects for many of the items; hence, there was a high propensity for item response distributions to be skewed towards the bottom category. Such response distribution, however, would be expected for a general population sample and was not deemed undesirable in this specific context. The items with the greatest floor effects (e.g., percentage in category 1 > 80%) were from the suicidality subdomain. Endorsement of these items would indicate the most severe levels of depression, so this pattern was consistent with the expected endorsement rate. In fact, only 10 of the 51 items did not display severe floor effects. These 10 items spanned four of the six subdomains (i.e., mood, cognition, behavior, somatic complaints), so it was likely that these items were indicators of milder depression levels as opposed to the items displaying floor effects that would be more appropriate (and necessary) for measuring moderate and severe depression levels. Keeping these limitations in mind (i.e., using data from a non-clinical sample), for the purpose of this paper, proposed strategies for item reduction focused on item redundancy as well as weak relationships of each item with the other items. Balancing these decisions with respective item content, we identified several candidate items for deletion that were largely consistent with those identified by Pilkonis et al. [11]. However, one should take into account that the reliability and validity of the final instrument is limited to a population sample, and thus cannot be readily applied to other – e.g. clinical – samples.

Second, the alpha reliability coefficient showed high internal consistency of all three instruments, i.e. the 51-, 28-, and 8-item version, respectively. However, these results need to be interpreted with care. As the reliability coefficient increases as the number of items increases, the interpretation of alpha of as many as 51 items – or 28 items – may not be useful as the value can be an artifact of the large number of items included. Nevertheless, it is reassuring that the 8-item short form showed a reliability of > 0.90, supporting the notion of good reliability of these PROMIS depression items [15]. One should keep in mind that good reliability can be achieved at the expense of item content, e.g. combining very similar but not necessarily complimentary items into one scale will result in highly reliable but not necessarily very valid scales. Thus, the development of an informative measurement scale requires a careful consideration of both reliability and validity.

Finally, the results of the confirmatory factor analyses showed satisfactory fit for all three instruments. Hence, there was sufficient support that the PROMIS depression items are indeed unidimensional. In other words, there was insufficient support for the alternative hypothesis, that is, that the items belonging to respective predefined subdomains (negative mood, decreases in positive emotions, cognitive deficits, negative self-image, and negative social cognition [7]) might be sufficiently different from items from other subdomains to justify a multidimensional depression construct. Nevertheless, the interested researcher could have considered alternative models, including multidimensional models or higher-order factor models, to investigate the tenability of such theoretical structures. In addition, more restricted models could have been considered to test specific substantive hypotheses. For example, one could test whether the individual factor loadings can be constrained to equality to test the tenability of the tau equivalence assumption. We chose to illustrate the application of a confirmatory unidimensional factor model with freely

estimated factor loadings to illustrate the potential of this type of analysis. The flexibility of the factor analytic framework can further be used to impose restrictions on (individual) model parameters to test (further) substantive hypotheses. One could, for example, investigate possible differences in the factor structure across groups of participants or across time (i.e., using the measurement invariance framework), and test possible differences in the underlying construct. However, illustration of the full potential of factor analytic techniques was outside the scope of the present paper.

There are a number of limitations to this CTT demonstration. First, PROMIS items were developed within a modern test theory framework, i.e. IRT; hence, these items were developed in a way that they are suitable to be administered as part of computerized-adaptive testing (CAT). Thus, item redundancy would not be observed in case of CAT, as only a subset of items would be administered. Also, any floor effects would be mitigated in practice, as the most severe items would be administered only to the most severe subjects. Therefore, the application of CTT methods in general to an instrument that was developed for and intended to be used with modern test theory methods is somewhat limited but applied here for illustrative purposes to compare three different test theory methods (CTT, IRT, RMT). Moreover, the illustration of CTT methods was limited in the sense that many more types of reliability and validity could have been considered. For example, alternative reliability tests include parallel-tests, split-half reliability, test-retest reliability, inter-rater reliability, etc. [17]. Internal consistency may also be investigated by inspecting inter-item correlational patterns or by using Structural Equation Modeling [44]. Additionally, construct validity may subsume convergent, discriminative, and predictive validity. Convergent and predictive validity are usually explored by investigating the associations between instrument scores and some gold standard. However, these types of gold standards are generally not available for PRO measures [45]. In the present case, the dataset that was distributed to the research teams lacked concurrent measures that could be used for assessing this type of validity and we also lacked a variable to identify subjects in the sample, for example, those with a clinical depression diagnosis; hence, these types of analyses were not possible and also beyond the scope of such paper. Finally, because this demonstration was conducted on an item bank rather than a static instrument, measurement properties were only explored at the item-level because the items are not intended to all be used to produce a score. When items can be combined to produce a domain score, CTT can be used to evaluate if the resulting scores are reliable, valid. Nevertheless, within the limits of a scientific paper aimed at giving an introduction to and illustration of a practical application of a test theory method such as CTT, it is hoped that this paper still succeeded in giving a sufficiently comprehensive overview of classical test theory methods.

Conclusions

The results of the psychometric analyses were considered against the conceptual framework as identified by the developers of the item bank. Overall, the items of the PROMIS Depression Item Bank performed well and seemed appropriate for measuring the unidimensional construct of depression. It is our hope that the illustration of CTT methods in this paper will help researchers to evaluate item- and scale-performance of PRO instruments. Among the advantages of CTT methods are that they are relatively easy to understand and apply, and that statistical software to perform the analyses is widely available. In conclusion, a combination of both classical and modern test theory methods will not only help evaluate the quality of a PRO instrument but may eventually also help advance the assessment and interpretation of psychological constructs that these instruments intend to measure.

Abbreviations

CAT: Computerized adaptive testing; CFA: Confirmatory factor analysis; CFI: Comparative fit index; CTT: Classical test theory; IRT: Item response theory; ISOQOL: International Society for Quality of Life Research; PRO: Patient-reported outcome; PROMIS: Patient-reported Outcomes Measurement Information System; RMSEA: Root mean square error of approximation; RMT: Rasch measurement theory; SIG: Special Interest Group; TLI: Tucker Lewis index; WLSMV: Weighted least squares means and variance adjusted

Acknowledgements

The authors would like to thank the Patient-Reported Outcomes Measurement Information System (PROMIS®) group for making available the data on the PROMIS Depression Item Bank used for illustrative purposes in this paper. The authors received writing and editorial support under the guidance of the authors from Drs. Chad Green and Amlan RayChaudhury (Clinical Outcomes Solutions, Chicago, USA).

Declarations

This paper was reviewed and endorsed by the International Society for Quality of Life Research (ISOQOL) Board of Directors as an ISOQOL publication and does not reflect an endorsement of the ISOQOL membership. All statements, findings and conclusions in this publication are solely those of the authors and do not necessarily represent the views of ISOQOL.

Authors' contributions

Conception or design of the work (SN, CC, SH, MV), data analysis and interpretation (SN, MV), manuscript preparation (SN, MV), critical revision (CC, SH) and final approval of the manuscript (SN, CC, SH, MV).

Funding

None.

Availability of data and materials

The datasets supporting the conclusions of this article are available from the PROMIS Health Organization, http://www.healthmeasures.net/explore-measurement-systems/promis

Ethics approval and consent to participate

Not applicable, use of publicly available data set.

Consent for publication

Not applicable.

Competing interests
CC reports personal fees from pharmaceutical companies, outside the submitted work. SH is the CEO of Clinical Outcomes Solutions, which received consulting fees from multiple pharmaceutical companies outside the submitted work. The other authors declare they have no competing interests.

Author details
[1]Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charité - Universitätsmedizin Berlin, Berlin, Germany. [2]Centre for Population Health Research, School of Health and Social Development, Deakin University, Geelong, Australia. [3]Outcometrix, 2912 NE Plaza Drive, Tucson, AZ 85716, USA. [4]Clinical Outcomes Solutions, 1820 East River Road, Suite 220, Tucson, AZ 85718, USA. [5]Department of Medical Psychology, Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands. [6]Department of Methods & Statistics Institute of Psychology, Leiden University, P.O. Box 9555, 2300, RB, Leiden, The Netherlands.

Received: 5 February 2019 Accepted: 6 June 2019
Published online: 30 July 2019

References

1. McDonald, R. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.
2. Spearman, C. (1904). "general intelligence," objectively determined and measured. *Am J Psychol, 15*(2), 201–292.
3. European Medicines Agency. Reflection paper on the regulatory guidance for the use of health-related quality of life (HRQL) measures in the evaluation of medicinal products. 2005 23 Mar 2016]; Available from: http://www.emea.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003637.pdf.
4. Food and Drug Administration. Guidance for industry. Patient-reported outcome measures: Use in medical product development to support labeling claims. 2009 23 Mar 2016]; Available from: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf.
5. Kluetz, P. G., et al. (2016). Focusing on Core patient-reported outcomes in Cancer clinical trials: Symptomatic adverse events, physical function, and disease-related symptoms. *Clin Cancer Res, 22*(7), 1553–1558.
6. Frank, L., Basch, E., & Selby, J. V. (2014). The PCORI perspective on patient-centered outcomes research. *Jama, 312*(15), 1513–1514.
7. Cella, D., et al. (2010). The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *J Clin Epidemiol, 63*(11), 1179–1194.
8. Alonso, J., et al. (2013). The case for an international patient-reported outcomes measurement information system (PROMIS(R)) initiative. *Health Qual Life Outcomes, 11*, 210.
9. DeVellis, R. F. (2006). Classical test theory. *Med Care, 44*(11 Suppl 3), S50–S59.
10. Mokkink, L. B., et al. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol, 63*(7), 737–745.
11. Pilkonis, P. A., et al. (2011). Item banks for measuring emotional distress from the patient-reported outcomes measurement information system (PROMIS(R)): Depression, anxiety, and anger. *Assessment, 18*(3), 263–283.
12. Radloff, L. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psy Measure, 1*(3), 385–401.
13. Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *J Gen Intern Med, 16*(9), 606–613.
14. Beck, A. T., et al. (1961). An inventory for measuring depression. *Arch Gen Psychiatry, 4*, 561–571.
15. Fayers, P., & Machin, D. (2016). *Quality of life: The assessment, analysis and reporting of patient-reported outcomes* (3rd ed.). Chichester: John Wiley & Sons, Ltd.
16. Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
17. Field, A. (2014). *Discovering statistics using IBM SPSS statistics* (4th ed.). London: SAGE Publications.
18. Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(8), 297–334.
19. Hair, J., et al. (2006). *Multivariate data analysis* (6th ed.). Upper Saddle River, NJ: Pearson Education, Inc.
20. Cortina, J. (1993). What is coefficient alpha? An examination of theory and applications. *J Appl Psychol, 78*(1), 98–104.
21. Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychol Assess, 8*(4), 350–353.
22. Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107–120.
23. Gadermann, A., Guhn, M., & Zumbo, B. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Pract Assess Res Eval, 17*(3), 1–13.
24. Frisbie, D. (1988). Reliability of scores from teacher-made tests. *Educ Meas Issues Pract, 7*(1), 25–35.
25. Bollen, K., & Arminger, G. (1991). Observational residuals in factor analysis and structural equation models. *Sociol Methodol, 21*, 235–262.
26. Bollen, K. (1989). *Structural equations with latent variables* (p. 514). New York: John Wiley & Sons.
27. Satorra, A., & Bentler, P. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. In *Proceedings of the business and economic statistics section of the American Statistical Association, 1988: P* (pp. 308–313).
28. Satorra, A., & Bentler, P. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In *Latent variable analysis: Applications for developmental research, A. von eye and C. Clogg* (pp. 399–419). Thousand Oaks, CA: Sage Publications.
29. Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychol Methods, 9*(4), 466–491.
30. Gerbing, D. and J. Anderson, Monte Carlo evaluations of goodness-of-fit indices for Structural Equation Models, in Testing Structural Equation Models, K. Bollen J. Long, 1993, Sage publications: Newbury Park, London, New Delhi.
31. Hu, L.-T. and P. Bentler, Evaluating model fit, in Structural Equation Modeling: concepts, issues, and applications, R. Hoyle, 1995, Sage Publications: Thousand Oaks. 76–99.
32. Browne, M., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivar Behav Res, 24*(4), 445–455.
33. Marsh, H., Balla, J., & McDonald, R. (1988). Goodness of fit indexes in confirmatory factor analysis: The effect of sample size. *Psychol Bull, 103*, 391–410.
34. Bentler, P. (1990). Comparative fit indexes in structural models. *Psychol Bull, 107*(2), 238–246.
35. Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation method, and model specification on structural equation modeling fit indexes. *Struct Equ Model, 6*, 56–83.
36. Hu, L., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct Equ Model, 6*(1), 1–55.
37. Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). *Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. MPR-online* (Vol. 8, pp. 23–74).
38. Marsh, H., Hau, K.-T., & Wen, Z. (2004). In search of Golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Struct Equ Model Multidiscip J, 11*(3), 320–341.
39. Jöreskog, K.G. and D. Sörbom, LISREL 8: User's reference guide. 2nd ed. 1996–2001, Lincolnwood, IL: Scientific Software International.
40. Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *J Stat Softw, 48*(2), 1–36.
41. Klem, M., et al. (2009). Building PROMIS item banks: Librarians as co-investigators. *Qual Life Res, 18*(7), 881–888.
42. Kelly, M. A., et al. (2011). Describing depression: Congruence between patient experiences and clinical assessments. *Br J Clin Psychol, 50*(1), 46–66.
43. Riley, W. T., et al. (2010). Patient-reported outcomes measurement information system (PROMIS) domain names and definitions revisions: Further evaluation of content validity in IRT-derived item banks. *Qual Life Res, 19*(9), 1311–1321.
44. Bentler, P. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika, 74*(1), 137–143.
45. Sloan, J. A., Cella, D., & Hays, R. D. (2005). Clinical significance of patient-reported questionnaire data: Another step toward consensus. *J Clin Epidemiol, 58*(12), 1217–1219.

Publisher's Note

2.3 Establishing the European Norm for the EORTC CAT Core

The EORTC CAT Core is one of the latest instruments developed by the EORTC QLG using the IRT framework (Petersen et al., 2018). Analogous to the EORTC QLQ-C30 that consists of 15 scales, the EORTC CAT Core consists of the same number of item banks, except for the overall quality of life scale of the EORTC QLQ-C30. A first international validation study of the EORTC CAT Core has already been published (Petersen et al., 2020). Despite the initial validation work on the EORTC CAT Core, however, an IRT-based PRO measure is only fully functional if the obtained scores can be linked to a reference population. Without such reference population, the IRT-based theta scores are placed on an arbitrary metric where the mean has little clinical or intuitive meaning. Hence, the transformation to a meaningful metric through the collection of general population norm data is crucial for the usefulness, comprehensibility and interpretability of the scores obtained via the EORTC CAT Core. Consequently, in parallel to the validation work, the EORTC QLG funded a study aimed at generating general population norm data for the EORTC CAT Core. This large international effort involved 14 collaborators and resulted in the creation of the European Norm for the EORTC CAT Core derived from a total of 11 countries of the European Union (EU). Following from this effort, the IRT-based scores generated from the EORTC CAT Core are now centred around a more meaningful mean, with a mean of 50 reflecting the mean of the European general population (Liegl et al., 2019).

The following abstract has been taken from the original peer-reviewed article:

Liegl G, Petersen MA, Groenvold M, Aaronson NK, Costantini A, Fayers PM, Holzner B, Johnson CD, Kemmler G, Tomaszewski KA, Waldmann A, Young TE, Rose M, Nolte S. Establishing the European Norm for the health-related quality of life domains of the computer-adaptive test EORTC CAT Core. *European Journal of Cancer*. 2019;107:133–41. DOI: https://doi.org/10.1016/j.ejca.2018.11.023

> *"OBJECTIVE: The computer-adaptive test (CAT) of the European Organisation for Research and Treatment of Cancer (EORTC), the EORTC CAT Core, assesses the same 15 domains as the EORTC QLQ-C30 health-related quality of life questionnaire but with increased precision, efficiency, measurement range and flexibility. CAT parameters for estimating scores have been established based on clinical data from cancer patients. This study aimed at establishing the European Norm for each CAT domain based on*

general population data. METHODS: We collected representative general population data across 11 European Union (EU) countries, Russia, Turkey, Canada and the United States (n >/= 1000/country; stratified by sex and age). We selected item subsets from each CAT domain for data collection (totalling 86 items). Differential item functioning (DIF) analyses were conducted to investigate cross-cultural measurement invariance. For each domain, means and standard deviations from the EU countries (weighted by country population, sex and age) were used to establish a T-metric with a European general population mean = 50 (standard deviation = 10). RESULTS: A total of 15,386 respondents completed the online survey (n = 11,343 from EU countries). EORTC CAT Core norm scores for all 15 countries were calculated. DIF had negligible impact on scoring. Domain-specific T-scores differed significantly across countries with small to medium effect sizes. CONCLUSION: This study establishes the official European Norm for the EORTC CAT Core. The European CAT Norm can be used globally and allows for meaningful interpretation of scores. Furthermore, CAT scores can be compared with sex- and age-adjusted norm scores at a national level within each of the 15 countries."

European Journal of Cancer 107 (2019) 133–141

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Original Research

Establishing the European Norm for the health-related quality of life domains of the computer-adaptive test EORTC CAT Core

G. Liegl [a], M.A. Petersen [b], M. Groenvold [b,c], N.K. Aaronson [d], A. Costantini [e], P.M. Fayers [f], B. Holzner [g], C.D. Johnson [h], G. Kemmler [g], K.A. Tomaszewski [i], A. Waldmann [j,k], T.E. Young [l], M. Rose [a,m], S. Nolte [a,n,*] on behalf of the EORTC Quality of Life Group

[a] *Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin Institute of Health, Berlin, Germany*
[b] *Department of Palliative Medicine, Bispebjerg Hospital, Copenhagen, Denmark*
[c] *Department of Public Health, University of Copenhagen, Copenhagen, Denmark*
[d] *Division of Psychosocial Research & Epidemiology, The Netherlands Cancer Institute, Amsterdam, the Netherlands*
[e] *Psycho-Oncology Unit, Sant'Andrea Hospital Sapienza, University of Rome, Rome, Italy*
[f] *Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, United Kingdom*
[g] *Department of Psychiatry, Psychotherapy and Psychosomatics, Innsbruck Medical University, Innsbruck, Austria*
[h] *University of Southampton, Southampton, United Kingdom*
[i] *Health Outcomes Research Unit, Department of Gerontology, Geriatrics, and Social Work, Faculty of Education, Ignatianum Academy, Krakow, Poland*
[j] *Institute of Social Medicine and Epidemiology, University of Luebeck, Luebeck, Germany*
[k] *Ministry for Health and Consumer Protection, Hamburg Cancer Registry, Hamburg, Germany*
[l] *East & North Hertfordshire NHS Trust, Mount Vernon Cancer Centre, Northwood, Middlesex, United Kingdom*
[m] *Quantitative Health Sciences, Outcomes Measurement Science, University of Massachusetts Medical School, Worcester, MA, USA*
[n] *Population Health Strategic Research Centre, School of Health and Social Development, Deakin University, Burwood, VIC, Australia*

Received 7 November 2018; accepted 10 November 2018
Available online 18 December 2018

DOI of original article: https://doi.org/10.1016/j.ejca.2018.11.024.
* *Corresponding author:* Sandra Nolte, Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charité – Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany.
E-mail address: sandra.nolte@charite.de (S. Nolte).

https://doi.org/10.1016/j.ejca.2018.11.023

KEYWORDS
Quality of life;
Computer-adaptive test;
Item response theory;
EORTC CAT Core;
Self-report;
Patient-reported outcomes;
General population;
Norm data;
Normative data;
Survey

Abstract ***Objective:*** The computer-adaptive test (CAT) of the European Organisation for Research and Treatment of Cancer (EORTC), the EORTC CAT Core, assesses the same 15 domains as the EORTC QLQ-C30 health-related quality of life questionnaire but with increased precision, efficiency, measurement range and flexibility. CAT parameters for estimating scores have been established based on clinical data from cancer patients. This study aimed at establishing the European Norm for each CAT domain based on general population data.

Methods: We collected representative general population data across 11 European Union (EU) countries, Russia, Turkey, Canada and the United States ($n \geq 1000$/country; stratified by sex and age). We selected item subsets from each CAT domain for data collection (totalling 86 items). Differential item functioning (DIF) analyses were conducted to investigate cross-cultural measurement invariance. For each domain, means and standard deviations from the EU countries (weighted by country population, sex and age) were used to establish a T-metric with a European general population mean = 50 (standard deviation = 10).

Results: A total of 15,386 respondents completed the online survey (n = 11,343 from EU countries). EORTC CAT Core norm scores for all 15 countries were calculated. DIF had negligible impact on scoring. Domain-specific T-scores differed significantly across countries with small to medium effect sizes.

Conclusion: This study establishes the official European Norm for the EORTC CAT Core. The European CAT Norm can be used globally and allows for meaningful interpretation of scores. Furthermore, CAT scores can be compared with sex- and age-adjusted norm scores at a national level within each of the 15 countries.

1. Introduction

The assessment of health-related quality of life (HRQoL) using patient-reported outcome (PRO) measures has become increasingly important to evaluate, monitor and improve the quality of cancer care [1–3]. One of the most frequently used PRO measures for cancer patients is the European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life core questionnaire QLQ-C30 [4,5]. It consists of 30 items covering five multi-item function scales (physical, role, social, cognitive and emotional function), three multi-item symptom scales (fatigue, nausea/vomiting and pain), six single-item scales assessing further aspects of HRQoL (dyspnoea, insomnia, appetite loss, constipation, diarrhoea and financial difficulties) and a scale on overall health/HRQoL.

The QLQ-C30 has been evaluated in many different cancer populations, and adequate psychometric properties were largely supported [6]. However, some studies found floor and ceiling effects [7,8]. Also, it is a well-known limitation of traditional HRQoL questionnaires, such as the QLQ-C30, that some questions might be irrelevant to the patient, which potentially increases respondent burden and results in less precise score estimates [9]. A promising solution to overcome such limitations is the use of computer-adaptive tests (CATs). CAT is methodologically based on item response theory (IRT). Using IRT, domain-specific item banks (i.e. lists of items measuring the same domain) can be calibrated on a common scale [9,10]. Fitting an IRT model provides item parameters that reflect the statistical relationship between an individual's response to a given item and his/her position on a domain scale. A major advantage of using IRT for scoring is that any item subset of an item bank can be used to estimate a person's domain score on the same continuous metric [11]. This enables CATs, in which HRQoL assessment is tailored to the individual, which increases measurement precision and range while reducing respondent burden, sample size requirements and study costs [10,12,13].

In 2006, the EORTC Quality of Life Group started developing the EORTC CAT Core, which is based on domain-specific item banks measuring the same dimensions as the QLQ-C30 [14,15]. In the item bank development process, different sources of information were collated, including literature reviews, qualitative input from various stakeholders and psychometric analyses of large international samples of cancer patients [14]. Item bank development for all domains was completed in 2016 [12,16–22]. To calibrate items of each bank, IRT models were estimated using data obtained from these clinical samples [22]. After item calibration, parameters are based on a z-score metric with a study population mean of 0 and standard deviation (SD) of 1. Such scores are 'arbitrary', hampering interpretation. Thus, the next step of item calibration is to link the CAT algorithm to general population norm data to simplify score interpretation. In this final step, it has become common practice to transform scores to a T-score metric with a general population mean of 50 (SD = 10).

The EORTC CAT Core development did not include this final step of transforming scores to T-scores. Therefore, for a more meaningful and sensible score interpretation, this study aimed at collecting representative data of the European general population to transform the current scoring to a T-score metric. Any score obtained via the EORTC CAT Core can then be interpreted in relation to this European mean. In addition, CAT norm scores are established for sex-, age- and country-specific subpopulations.

2. Methods

2.1. Sampling

To collect general population data to establish the 'European CAT Norm', we subcontracted GfK SE (http://www.gfk.com/). Data were collected via online surveys in March/April 2017 in 11 European Union (EU) countries (Austria, Denmark, France, Germany, Hungary, Italy, the Netherlands, Poland, Spain, Sweden and the United Kingdom), Russia, Turkey, Canada and the United States for comparative purposes. We stratified data collection by sex and age groups (18–39, 40–49, 50–59, 60–69, 70+ years), with a target sample size of each sex × age × country subgroup of n = 100, leading to an anticipated sample size of n = 1000/country. Assuming a T-scale with a mean = 50 (SD = 10), this sample size allows for estimating the population T-score mean of each country (n = 1000) within ±0.6 T-scores (95% confidence interval), which was considered to be sufficiently precise. Moreover, this sampling design is sufficient to investigate differential item functioning (DIF) using logistic regression analysis [23].

2.2. Selection of items

The full item pool of the EORTC CAT Core consists of 14 item banks for the function and symptom-related HRQoL domains in the QLQ-C30 plus two global items forming the 15th scale for *overall HRQoL*. The number of items per bank ranges between 7 (*appetite loss*) and 34 items (*fatigue, cognitive function*). In total, 260 items are available for CAT assessment. For establishing CAT norm values, 86 items were selected, consisting of all QLQ-C30 items plus four additional items per domain (56 items). The selection of the 56 items was based on high measurement precision and adequate range of measurement as indicated by the items' psychometric properties and content validity considerations, that is, all aspects of a given domain had to be covered.

2.3. Differential item functioning

DIF analyses are used to evaluate whether items measure the same underlying construct in different subpopulations [24], a central requirement for establishing a common norm across subpopulations. We investigated DIF regarding country, sex and age groups using ordinal logistic regression [23,25]. A regression was modelled for each item, with the item response as the dependent variable and the IRT-based domain score as the independent variable. If adding the grouping variable of interest (country/sex/age) to this model as an independent variable leads to a change in the Nagelkerke R^2 coefficient $\geq$.035, this indicates potentially relevant DIF [20,26]. If DIF was identified, we evaluated its practical importance by calculating the standardised mean difference (SMD) between scores derived from all available items of a domain versus scores derived from a reduced item set, that is, excluding items showing DIF. If SMD was $\leq$.2 (small effect size [ES] [27]), the practical impact of DIF was considered to be negligible and affected items were kept.

2.4. Establishing the European CAT norm

For establishing the 'European CAT Norm', we used general population data from the 11 EU countries. In a first step, we scored the data based on the previously established 'arbitrary' IRT-based z-score metric and calculated means/SDs for each CAT domain. To correct for over- or under-representation of subgroups, we weighted scores by country population size, sex and age distribution, with the youngest age group further divided into 18–29 and 30–39 years. Individual weighting factors were calculated for each country × sex × age group based on general population distribution statistics for 2015 [28] using the formula:

Weighting factor = percentage of subgroup in population/percentage of subgroup in sample.

After estimating weighted means and SDs, scores were transformed to a T-score metric using linear transformation to establish the 'European CAT Norm' with mean = 50 and SD = 10 using the formula:

$$\text{T-score} = 10 * (\text{z-score} - \text{z-score mean})/\text{z-score SD} + 50.$$

Using these formulas, we calculated European norm scores (means, SDs) for each CAT domain overall and by sex and age. Furthermore, to establish norm scores for each of the 15 countries, national T-score means/SDs were calculated using country-specific sex- and age-weighting factors.

2.5. Determining the extent of subgroup differences

For each CAT domain, we investigated T-score differences between countries, sex, age groups and dichotomised educational levels (less than postcompulsory education and at least some postcompulsory education). We conducted covariance analyses for each independent

variable, entering the remaining three variables as covariates. When used as a covariate, *age* was entered as continuous variable. Statistical significance was implied by P value <.01. We interpreted partial eta^2 values of .01 ($\approx R^2$ = 1%), .06 ($\approx R^2$ = 6%) and .14 ($\approx R^2$ = 14%) as small, medium and large ES, respectively [27].

For analysing DIF, R 3.1.2 was applied using the package *lordif* version 0.3-3 [25,29]. For all other statistical analyses, we used IBM SPSS Statistics®, version 22 [30].

3. Results

3.1. Sample

The total sample size was N = 15,386 for the full sample and n = 11,343 for the EU sample (Table 1). Further details on sampling and sociodemographic data are provided elsewhere [35].

3.2. Psychometric properties

DIF by country and age was detected in two items of the *physical function* (PF) scale. As the ES of excluding versus including the DIF-items when estimating PF scores in the full sample were very small (SMD = .01), all items of the PF scale were retained for further analyses. For the Hungarian data, one item of the *sleep problems* scale (SL4) had to be excluded from further analyses due to a translation error.

3.3. European CAT norm scores

Table 2 presents domain-specific *z-score* distributions in the EU countries, weighted by country population size, sex and age. In all CAT domains, except for *dyspnoea*, all mean scores indicate better HRQoL in the general population compared with scores in the original cancer populations (which by way of model estimation has mean = 0).

In Table 3, the final European CAT Norm *T-scores* (means and SDs) for each domain are reported for the EU sample overall (by definition with mean = 50) and by sex and age. Covariance analyses indicated higher HRQoL in men than in women (P < .001) in all domains but ES were small, ranging from eta^2 = .001 for *diarrhoea* and *nausea/vomiting* to eta^2 = .020 and .022 for *emotional function* and *PF*, respectively. Age also had a statistically significant effect on HRQoL (P < .001 for each domain); however, the relational patterns were inconsistent across CAT scales, and ES were small for most domains, ranging from eta^2 = .002 for *pain* to eta^2 = .020 for *fatigue*. In three domains, a larger and relatively linear age effect was found: *emotional function* (eta^2 = .047) and *nausea/vomiting* (eta^2 = .034) improved while *PF* scores (eta^2 = .066) worsened with age.

Table 1
EORTC CAT Core general population norm data sample characteristics.

Sociodemographic variable	Full sample (15 countries; N = 15,386)	Norm sample (11 EU countries; n = 11,343)
	n (%)	n (%)
Age, years		
18–29	1177 (7.6)	883 (7.8)
30–39	1902 (12.4)	1370 (12.1)
40–49	3049 (19.8)	2248 (19.8)
50–59	3059 (19.9)	2253 (19.9)
60–69	3138 (20.4)	2337 (20.6)
70+	3061 (19.9)	2252 (19.9)
Sex		
Female	7650 (49.7)	5623 (49.6)
Male	7736 (50.3)	5720 (50.4)
Education		
Less than compulsory education	183 (1.2)	95 (.8)
Compulsory (left school at the minimum school leaving age)	1509 (9.8)	897 (7.9)
Some postcompulsory (some school after reaching school leaving age without reaching university entrance qualifications [e.g. A levels])	2050 (13.3)	1954 (17.2)
Postcompulsory below university (e.g. reaching A levels)	4405 (28.6)	3408 (30.0)
University degree (bachelor's or equivalent level)	3716 (24.2)	2689 (23.7)
Postgraduate degree (master's, doctorate or equivalent level)	3337 (21.7)	2131 (18.8)
Prefer not to answer	186 (1.2)	169 (1.5)
Country		
Austria	1002 (6.5)	1002 (8.8)
Denmark	1003 (6.5)	1003 (8.8)
France	1001 (6.5)	1001 (8.8)
Germany	1006 (6.5)	1006 (8.9)
Hungary	1053 (6.8)	1053 (9.3)
Italy	1036 (6.7)	1036 (9.1)
The Netherlands	1000 (6.5)	1000 (8.8)
Poland	1024 (6.7)	1024 (9.0)
Spain	1165 (7.6)	1165 (10.3)
Sweden	1027(6.7)	1027 (9.1)
The United Kingdom	1026 (6.7)	1026 (9.0)
Russia	1007 (6.5)	–
Turkey	1023 (6.6)	–
Canada	1004 (6.5)	–
The United States	1009 (6.6)	–

Except for *diarrhoea*, scores were also significantly associated with educational level (P < .01) with higher educated individuals reporting better HRQoL scores (data not shown). However, these ES were very small (all eta^2 ≤ .015).

Table 4 presents CAT T-scores for all 15 countries. Within EU countries, domain scores in Poland indicated relatively low HRQoL, while scores were comparatively high in Austria and the Netherlands. In the United States and Canada, score distributions were relatively

Table 2
z-score distribution in the EU countries (n = 11,343) of the EORTC CAT Core domain scales.

EORTC CAT Core domain	Number of items	Mean	SD	Median	Minimum	Maximum	Skewness
Physical function	9	**.2327**	**.81453**	.1960	−2.69	1.51	−.312
Role function	6	**.4491**	**.82772**	.6990	−2.19	1.18	−.912
Emotional function	8	**.0813**	**.98832**	.0982	−2.76	1.38	−.394
Cognitive function	6	**.2466**	**.97733**	.3490	−2.98	1.24	−.777
Social function	6	**.3203**	**.93293**	1.0150	−2.62	1.02	−1.097
Fatigue	7	**−.1346**	**.92938**	−.1080	−1.45	2.42	.375
Nausea/vomiting	6	**−1.8961**	**.78304**	−2.2130	−2.21	2.14	2.748
Pain	6	**−.8193**	**1.20929**	−.9480	−2.03	2.43	.612
Dyspnoea	5	**.1787**	**.94265**	.2050	−.68	2.91	.674
Sleep problems	5	**−.3120**	**1.08615**	−.2320	−1.68	2.36	.461
Appetite loss	5	**−.2141**	**.77220**	−.6760	−.68	2.61	1.453
Constipation	5	**−.4222**	**.74434**	−.6020	−1.07	2.21	.928
Diarrhoea	5	**−.4564**	**.67354**	−.8460	−.85	2.20	1.535
Financial difficulties	5	**−.3617**	**.78364**	−.8310	−.83	2.38	1.497
Overall HRQoL	2	**.1163**	**.93597**	.0390	−2.58	1.91	−.089

EORTC CAT = The computer-adaptive test of the European Organisation for Research and Treatment of Cancer; EU = European Union; HRQoL = health-related quality of life. Bold values = The z-score mean and SD values were used in the formula described in section 2.4 for transforming the z-scores to T-scores.

similar to the EU countries. In contrast, mean scores of the Russian and Turkish general populations indicated worse HRQoL in most CAT domains compared to the EU average. In the covariance analyses, T-scores differed significantly across countries in all CAT domains ($P < .001$). However, ES were small for each domain ($eta^2 < .06$).

4. Discussion

The EORTC CAT Core is the first disease-specific computer-adaptive PRO assessment system developed across different countries for measuring a wide range of HRQoL aspects relevant to cancer patients. In an extensive development and psychometric evaluation process, the EORTC CAT has been proven to be a more precise, efficient and flexible measurement instrument compared to the traditional QLQ-C30 static questionnaire [22].

This study established the official 'European CAT Norm' based on general population data from 11 EU countries for a more meaningful and sensible interpretation of EORTC CAT scores. The domain-specific means and SDs presented herein are now implemented in the EORTC CAT scoring algorithm using a standardised scale centred to the European general population with a mean of 50 (SD = 10). Additionally, we present norm scores per country and for sex- and age-specific subgroups. This allows for a meaningful and detailed interpretation of cancer patients' scores.

Similar to our findings presented in the EORTC QLQ-C30 norm data paper [35], some group differences were observed. For example, men tended to score somewhat better than women, which is consistent with other QLQ-C30 norm data studies [31]. Furthermore, some observed age differences were counterintuitive, with the youngest participants showing lowest/worst scores in some function scales, which has also been observed by others in the application of item banks [32]. Due to these group differences, we recommend the use of sex- and age-matched norm data for the most sensible and meaningful score interpretation of data from cancer patients obtained via the EORTC CAT Core.

The observed differences between countries need to be taken at face value. It is conceivable that these differences reflect 'true' differences in HRQoL; however, it is also possible that some of these differences either reflect differences due to slightly different meanings between language versions or they reflect cultural differences, for example, in terms of culture-related health perceptions, expectations or response styles. Given the vast experience with questionnaire translation and cultural adaptation of items at the EORTC headquarters and findings in the literature showing language-related DIF to be negligible [33], we assume the observed differences to be 'true' country differences in HRQoL until further evidence is found that supports or refutes our hypothesis. Furthermore, our tests of country-DIF as presented herein show minimal impact of DIF, providing sufficient support for our assumption of true intercountry differences.

Our study has some limitations. First, it is not clear whether online panels are truly representative of the general population despite panel research companies claiming they are. As an increasingly large proportion of people have access to the Internet, the potential problem of representativeness is getting smaller but still remains, especially in countries such as Turkey where GfK had to carry out telephone interviews to achieve sampling quotas (for details see [35]). Our data suggest representativeness regarding most sample characteristics except for educational status; however, when testing for the influence of educational level, we found that the practical consequences were negligible. These findings

Table 3
European norm T-scores (based on 11 EU countries) for each EORTC CAT Core domain: mean scores (*M*) and standard deviations (SDs) by sex and age groups[a].

Domain	Full sample, *M* (SD)			18–29 years, *M* (SD)			30–39 years, *M* (SD)			40–49 years, *M* (SD)			50–59 years, *M* (SD)			60–69 years, *M* (SD)			70+ years, *M* (SD)		
	All	Male	Female	All	Male	Female	All	Male	Female	All	Male	Female	All	Male	Female	All	Male	Female	All	Male	Female
Physical function	50.00	51.52	48.56	52.83	53.75	51.86	52.54	53.82	51.23	51.47	52.95	49.99	49.34	50.89	47.83	47.40	48.60	46.30	45.36	47.24	44.02
	(10.00)	(10.23)	(9.56)	(9.53)	(10.38)	(8.44)	(9.83)	(9.93)	(9.55)	(10.00)	(9.93)	(9.87)	(9.74)	(9.87)	(9.38)	(9.26)	(9.60)	(8.79)	(9.29)	(9.62)	(8.81)
Role function	50.00	50.44	49.59	51.20	50.12	52.32	50.91	51.16	50.66	50.78	51.26	50.29	49.78	50.59	48.99	49.48	50.13	48.89	47.47	49.16	46.27
	(10.00)	(9.83)	(10.14)	(9.74)	(10.56)	(8.66)	(9.73)	(9.54)	(9.91)	(10.09)	(9.84)	(10.32)	(10.17)	(9.81)	(10.45)	(9.96)	(9.50)	(10.34)	(9.86)	(9.18)	(10.14)
Emotional function	50.00	51.19	48.88	48.50	49.86	47.09	48.21	48.98	47.42	49.19	50.40	47.96	49.71	51.05	48.40	52.16	53.40	51.02	52.92	54.81	51.57
	(10.00)	(9.84)	(10.02)	(10.84)	(10.93)	(10.55)	(10.36)	(10.36)	(10.32)	(10.15)	(9.71)	(10.43)	(9.60)	(9.54)	(9.49)	(9.08)	(8.65)	(9.33)	(8.48)	(7.46)	(8.90)
Cognitive function	50.00	50.65	49.39	48.33	48.73	47.91	49.57	50.24	48.89	49.82	50.78	48.85	50.26	51.40	49.14	51.56	51.84	51.31	51.11	51.80	50.61
	(10.00)	(10.04)	(9.93)	(11.11)	(11.57)	(10.61)	(10.93)	(10.83)	(10.99)	(10.35)	(10.08)	(10.53)	(9.78)	(9.66)	(9.78)	(8.45)	(8.41)	(8.49)	(8.15)	(7.74)	(8.40)
Social function	50.00	50.38	49.64	49.70	49.74	49.66	49.10	49.41	48.77	48.98	49.46	48.51	49.75	50.62	48.91	51.19	51.68	50.73	51.56	52.08	51.19
	(10.00)	(9.84)	(10.14)	(10.44)	(10.58)	(10.30)	(10.60)	(10.53)	(10.67)	(10.55)	(10.20)	(10.87)	(10.06)	(9.67)	(10.36)	(9.21)	(8.84)	(9.53)	(8.44)	(7.95)	(8.75)
Fatigue	50.00	48.84	51.10	50.89	49.96	51.87	51.18	50.06	52.33	50.67	49.21	52.15	50.09	48.88	51.28	48.31	47.42	49.13	48.38	46.54	49.69
	(10.00)	(9.81)	(10.06)	(9.61)	(9.71)	(9.41)	(10.15)	(9.67)	(10.50)	(10.35)	(9.89)	(10.60)	(10.06)	(9.98)	(10.01)	(9.84)	(9.63)	(9.97)	(9.64)	(9.46)	(9.56)
Nausea/vomiting	50.00	49.96	50.04	52.11	53.27	50.90	52.11	52.28	51.94	50.06	49.42	50.71	49.07	48.48	49.65	48.16	47.57	48.70	47.79	47.00	48.36
	(10.00)	(10.27)	(9.74)	(12.70)	(14.14)	(10.86)	(11.88)	(12.40)	(11.33)	(9.97)	(9.16)	(10.69)	(8.31)	(7.59)	(8.93)	(7.09)	(6.00)	(7.93)	(6.36)	(4.33)	(7.43)
Pain	50.00	49.35	50.61	49.17	49.25	49.09	49.87	49.79	49.95	50.06	49.41	50.71	50.80	49.79	51.80	49.87	49.30	50.39	50.39	48.44	51.77
	(10.00)	(9.74)	(10.20)	(9.83)	(10.01)	(9.63)	(10.15)	(9.96)	(10.35)	(10.21)	(9.84)	(10.52)	(10.15)	(9.85)	(10.35)	(9.92)	(9.42)	(10.33)	(9.69)	(9.06)	(9.88)
Dyspnoea	50.00	49.54	50.44	49.12	49.52	48.70	49.50	49.00	50.01	49.18	48.50	49.88	50.01	49.40	50.62	50.53	50.24	50.80	51.98	51.00	52.68
	(10.00)	(10.01)	(9.97)	(9.67)	(10.39)	(8.84)	(9.98)	(10.09)	(9.85)	(9.87)	(9.55)	(10.14)	(9.90)	(9.69)	(10.08)	(9.98)	(9.90)	(10.05)	(10.38)	(10.18)	(10.47)
Sleep problems	50.00	48.97	50.97	48.52	47.40	49.68	50.31	49.77	50.86	50.56	49.76	51.36	51.27	49.70	52.81	49.89	48.88	50.83	49.75	48.64	50.54
	(10.00)	(9.82)	(10.08)	(9.92)	(9.86)	(9.85)	(10.25)	(10.20)	(10.28)	(10.23)	(9.95)	(10.45)	(10.45)	(10.14)	(10.52)	(9.64)	(9.31)	(9.85)	(9.17)	(8.86)	(9.31)
Appetite loss	50.00	49.69	50.30	51.95	52.52	51.34	51.71	51.63	51.78	49.84	48.96	50.73	49.19	48.52	49.85	48.41	47.83	48.95	48.28	47.24	49.02
	(10.00)	(9.84)	(10.14)	(11.24)	(11.78)	(10.61)	(11.05)	(10.91)	(11.20)	(9.85)	(8.92)	(10.64)	(9.23)	(8.85)	(9.55)	(8.72)	(7.98)	(9.33)	(8.47)	(7.53)	(9.01)
Constipation	50.00	49.62	50.36	51.40	51.34	51.46	51.44	50.69	52.22	49.72	49.03	50.42	49.41	48.56	50.24	48.59	48.38	48.79	48.97	48.98	48.95
	(10.00)	(9.70)	(10.26)	(10.69)	(10.95)	(10.41)	(10.78)	(10.51)	(11.01)	(10.01)	(9.27)	(10.66)	(9.50)	(8.93)	(9.97)	(9.01)	(8.62)	(9.35)	(9.23)	(8.64)	(9.62)
Diarrhoea	50.00	50.53	49.50	51.21	52.64	49.71	52.01	52.70	51.31	50.15	50.72	49.56	49.58	49.60	49.55	48.22	48.41	48.05	48.36	47.74	48.80
	(10.00)	(10.34)	(9.64)	(11.12)	(12.00)	(9.90)	(11.11)	(11.39)	(10.78)	(10.10)	(10.40)	(9.76)	(9.59)	(9.73)	(9.46)	(8.21)	(8.14)	(8.27)	(8.41)	(7.22)	(9.14)
Financial difficulties	50.00	49.90	50.10	50.25	51.32	49.13	51.07	50.96	51.20	50.83	50.22	51.45	50.24	49.45	51.01	48.92	48.56	49.24	48.45	47.92	48.82
	(10.00)	(9.84)	(10.15)	(10.52)	(10.96)	(9.92)	(10.86)	(10.78)	(10.94)	(10.66)	(10.14)	(11.13)	(9.86)	(9.19)	(10.42)	(9.07)	(8.71)	(9.38)	(8.25)	(7.51)	(8.72)
Overall HRQoL	50.00	50.83	49.21	51.11	52.28	49.89	49.66	50.46	48.85	49.26	49.95	48.55	49.13	49.84	48.44	50.17	50.44	49.92	50.46	51.77	49.53
	(10.00)	(10.01)	(9.93)	(10.05)	(10.58)	(9.32)	(9.87)	(9.82)	(9.86)	(9.96)	(9.57)	(10.30)	(10.33)	(10.35)	(10.26)	(9.96)	(9.59)	(10.29)	(9.64)	(9.55)	(9.60)

EORTC CAT = The computer-adaptive test of the European Organisation for Research and Treatment of Cancer; EU = European Union; HRQoL = health-related quality of life.
[a] The European general population has a mean T-score of 50 (SD = 10).

Table 4
Country-specific EORTC CAT Core T-score[a] distributions.

Domain	AUT, *M* (SD)	CAN, *M* (SD)	DNK, *M* (SD)	FRA, *M* (SD)	DEU, *M* (SD)	HUN, *M* (SD)	ITA, *M* (SD)	NLD, *M* (SD)	POL, *M* (SD)	RUS, *M* (SD)	ESP, *M* (SD)	SWE, *M* (SD)	TUR, *M* (SD)	GBR, *M* (SD)	USA, *M* (SD)	Partial η^2
Physical function	52.31 (9.20)	50.19 (10.02)	50.46 (10.01)	51.24 (9.71)	48.33 (10.64)	51.30 (8.84)	50.37 (8.61)	52.74 (9.54)	48.34 (8.35)	44.69 (6.85)	51.35 (9.42)	52.33 (9.16)	46.76 (6.77)	49.02 (11.71)	49.18 (12.66)	.054
Role function	52.91 (8.39)	50.04 (9.91)	49.55 (9.99)	51.42 (9.35)	49.13 (10.76)	51.66 (8.25)	50.59 (9.55)	52.52 (9.30)	48.33 (9.12)	47.35 (8.51)	50.02 (9.14)	52.31 (8.45)	47.30 (8.41)	48.57 (11.33)	49.14 (11.49)	.029
Emotional function	52.35 (8.97)	50.76 (9.53)	51.95 (10.62)	51.62 (9.89)	50.63 (9.92)	49.67 (8.89)	48.49 (9.22)	53.47 (9.31)	47.27 (9.80)	47.21 (8.83)	50.77 (9.53)	51.14 (9.08)	45.76 (9.33)	48.60 (11.07)	50.50 (11.08)	.040
Cognitive function	52.13 (8.71)	49.62 (9.79)	49.66 (10.31)	51.14 (9.35)	50.27 (10.34)	49.75 (9.02)	50.33 (9.38)	52.83 (8.75)	47.90 (10.02)	46.51 (8.79)	50.29 (9.38)	50.77 (9.02)	45.58 (10.05)	48.19 (11.31)	48.67 (11.61)	.028
Social function	53.05 (7.81)	49.54 (10.67)	50.00 (10.46)	51.80 (9.03)	49.93 (10.07)	51.25 (8.42)	50.35 (9.20)	52.89 (8.27)	46.63 (10.53)	47.67 (9.77)	50.77 (9.65)	52.48 (8.46)	47.62 (10.17)	47.72 (11.44)	48.22 (11.57)	.033
Fatigue	48.39 (9.35)	50.03 (9.28)	50.46 (10.57)	48.77 (10.20)	51.17 (10.54)	49.99 (8.71)	49.22 (9.45)	47.48 (9.49)	52.44 (8.20)	53.79 (8.69)	48.30 (9.49)	48.91 (8.82)	53.65 (8.79)	51.18 (10.79)	50.86 (10.93)	.035
Nausea/vomiting	47.44 (5.77)	50.30 (9.95)	51.35 (10.88)	48.87 (8.76)	49.96 (10.34)	49.43 (8.72)	50.79 (10.34)	48.13 (6.89)	51.11 (10.98)	51.42 (10.11)	49.49 (9.34)	48.83 (7.66)	54.32 (12.08)	51.10 (11.39)	52.51 (12.36)	.018
Pain	48.36 (8.98)	50.38 (9.74)	50.25 (9.97)	48.64 (9.45)	50.84 (10.91)	50.21 (8.83)	49.06 (9.20)	47.48 (8.79)	51.95 (9.19)	52.10 (9.27)	50.31 (9.35)	48.71 (9.32)	52.03 (8.83)	50.73 (11.15)	51.27 (10.96)	.016
Dyspnoea	46.91 (8.69)	49.98 (9.93)	48.46 (9.54)	48.95 (9.70)	50.72 (10.89)	48.33 (8.56)	51.59 (9.65)	47.87 (8.80)	50.08 (9.32)	52.95 (9.32)	48.66 (9.19)	50.72 (7.73)	52.54 (9.41)	50.77 (10.89)	51.25 (11.29)	.026
Sleep problems	48.70 (9.67)	51.62 (10.01)	50.47 (9.94)	49.22 (9.99)	50.96 (10.99)	48.72 (8.78)	48.25 (8.81)	48.40 (8.87)	50.57 (9.47)	51.47 (9.45)	49.23 (9.19)	48.80 (8.64)	51.82 (8.71)	52.33 (10.68)	51.46 (10.66)	.017
Appetite loss	46.87 (7.20)	50.25 (10.20)	51.09 (10.87)	49.06 (9.28)	49.92 (10.26)	49.12 (8.69)	49.51 (9.26)	47.36 (7.50)	51.78 (10.76)	51.75 (10.04)	49.45 (9.52)	49.17 (9.00)	56.34 (11.09)	52.11 (11.40)	51.84 (11.35)	.035
Constipation	47.27 (8.26)	50.35 (10.14)	49.79 (10.16)	49.24 (9.35)	48.96 (10.05)	50.74 (9.04)	50.49 (9.80)	46.45 (7.56)	52.82 (10.63)	53.16 (9.75)	51.15 (9.96)	47.77 (9.04)	55.19 (11.07)	50.65 (10.74)	51.88 (11.75)	.045
Diarrhoea	48.43 (8.52)	51.32 (10.40)	50.51 (10.15)	48.63 (8.86)	49.87 (10.33)	50.14 (9.62)	50.25 (10.02)	48.04 (8.28)	51.99 (11.05)	52.09 (10.88)	49.59 (9.41)	48.83 (8.44)	53.57 (11.09)	51.22 (10.82)	52.06 (11.57)	.016
Financial difficulties	47.12 (7.66)	50.46 (10.96)	50.41 (10.44)	47.99 (8.63)	50.04 (10.23)	51.62 (10.47)	50.15 (9.64)	46.97 (7.10)	53.38 (10.81)	54.32 (11.22)	49.73 (9.35)	47.22 (7.58)	54.93 (11.05)	51.25 (11.41)	52.60 (12.23)	.055
Overall HRQoL	54.65 (9.78)	49.89 (9.54)	50.81 (11.01)	50.74 (9.27)	50.49 (10.04)	49.82 (9.40)	49.13 (9.32)	55.79 (9.81)	47.10 (9.11)	46.91 (8.79)	50.34 (9.89)	51.74 (10.47)	47.33 (10.22)	48.50 (10.80)	49.19 (10.49)	.059

Country codes: AUT = Austria; CAN=Canada; DEU = Germany; DNK = Denmark; FRA = France; HUN = Hungary; ITA = Italy; NLD = the Netherlands; POL = Poland; RUS = Russia; ESP = Spain; SWE=Sweden; TUR = Turkey; GBR = United Kingdom, USA = United Stated of America.
EORTC CAT = The computer-adaptive test of the European Organisation for Research and Treatment of Cancer; HRQoL = health-related quality of life.
[a] T-scores showed statistically significant differences ($P < .001$) between countries in each EORTC CAT Core domain.

support the notion that our data are suitable to establish the 'European CAT Norm' as well as valid norm scores for the 15 countries included in our study. Second, using online panels, we were able to collect a large database of N = 15,386 covering 15 countries and balanced by sex and age groups from 18 to 70+ years. This large sample size enabled detailed DIF analyses and precise T-score transformations. Of note, using linear transformation to transform z-scores to T-scores based on general population data, in which a substantial proportion of participants have 'perfect' scores (e.g. no pain), leads to distributional properties of the T-scores that do not follow a normal distribution. Linear transformation into T-scores is the current standard and also used by, for example, Patient-Reported Outcomes Measurement Information System (PROMIS) [34]. However, these specific distributional properties of the T-scores need to be kept in mind when interpreting the scores. Finally, for practical reasons, we were only able to collect data on item subsets from each EORTC CAT Core item bank. It was not feasible to collect data on 262 items as this would bring other problems such as respondent burden. We have to assume that selected items are representative for the full item banks. As the included items were carefully selected based on each item's psychometric properties and content validity considerations, the data presented herein are robust, state-of-the-art general population norm data for the EORTC CAT Core.

5. Conclusions

In this article, we present representative general population data for the cancer-specific computer-adaptive PRO assessment system EORTC CAT Core across 11 EU countries, Russia, Turkey, Canada and the United States. By defining the 'European CAT Norm', that is, a common European Norm for the EORTC CAT Core, scores from cancer patients obtained via this new instrument can be easily interpreted. In addition, EORTC CAT Core norm scores are provided for age-, sex- and country-specific subpopulations in 15 countries allowing for meaningful score interpretation and comparisons across countries and cultures.

Conflict of interest statement

None declared.

Ethical statement

Ethical approval was not sought as this study is solely based on panel research data. As opposed to medical research where medical professional codes of conduct apply, there is widespread agreement that health research involving volunteers from the general population is not subject to ethical approval. Both the European Pharmaceutical Market Research Association (EphMRA) and the NHS Health Research Authority specify that this type of research does not require ethical approval as long as the research conforms to ethical guidelines. Our online survey was carried out by the panel research company GfK SE, which is member of EphMRA. The multinational survey conformed to the required ethical standards by obtaining informed consent from all participants and collecting data completely anonymously. Any identification of the respondents through the authors is impossible.

Acknowledgements

This research was funded by the European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Group (grant number 001 2015). The work of Bernhard Holzner was partially funded by the Austrian Science Fund (FWF #P26930). The authors thank the many translators who kindly volunteered their time to embark on the task of standardising the sociodemographic variables across languages and cultures. In addition to the collaborators of this project, special thanks go to Susan Bartlett for making the variables in English and French suitable for the Canadian context; Eveline Bleiker, Jacobien Kieffer, Marieke van Leeuwen for Dutch; Thierry Conroy for French; Agnes Czimbalmos for Hungarian; Alice Iuso for Italian; TatyanaIonova for Russian; Juan Ignacio Arraras for Spanish; Eva Hammerlid, Yvonne Brandberg for Swedish; Deniz Yüce for Turkish; and Claire Snyder for English for the U.S.American context.

References

[1] Blazeby JM, Avery K, Sprangers M, Pikhart H, Fayers P, Donovan J. Health-related quality of life measurement in randomized clinical trials in surgical oncology. J Clin Oncol 2006;24:3178–86.

[2] Boele FW, Douw L, Reijneveld JC, Robben R, Taphoorn MJB, Aaronson NK, et al. Health-related quality of life in stable, long-term survivors of low-grade glioma. J Clin Oncol 2015;33:1023–9.

[3] Cella D, Grünwald V, Nathan P, Doan J, Dastani H, Taylor F, et al. Quality of life in patients with advanced renal cell carcinoma given nivolumab versus everolimus in CheckMate 025: a randomised, open-label, phase 3 trial. Lancet Oncol 2016;17:994–1003.

[4] Fayers P, Bottomley A. Quality of life research within the EORTC—the EORTC QLQ-C30. Eur J Cancer 2002;38:125–33.

[5] Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993;85:365–76.

[6] Luckett T, King MT, Butow PN, Oguchi M, Rankin N, Price MA, et al. Choosing between the EORTC QLQ-C30 and FACT-G for measuring health-related quality of life in cancer clinical research: issues, evidence and recommendations. Ann Oncol 2011;22:2179–90.

[7] Holzner B, Bode RK, Hahn EA, Cella D, Kopp M, Sperner-Unterweger B, et al. Equating EORTC QLQ-C30 and FACT-G

scores and its use in oncological research. Eur J Cancer 2006;42:3169–77.
[8] Knobel H, Loge JH, Brenne E, Fayers P, Hjermstad MJ, Kaasa S. The validity of EORTC QLQ-C30 fatigue scale in advanced cancer patients and cancer survivors. Palliat Med 2003;17:664–72.
[9] Cella D, Gershon R, Lai J-S, Choi S. The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. Qual Life Res 2007;16:133–41.
[10] Bjorner JB, Chang C-H, Thissen D, Reeve BB. Developing tailored instruments: item banking and computerized adaptive assessment. Qual Life Res 2007;16:95–108.
[11] Liegl G, Gandek B, Fischer HF, Bjorner JB, Ware JE, Rose M, et al. Varying the item format improved the range of measurement in patient-reported outcome measures assessing physical function. Arthritis Res Ther 2017;19:66.
[12] Petersen MA, Aaronson NK, Arraras JI, Chie WC, Conroy T, Costantini A. The EORTC computer-adaptive tests measuring physical functioning and fatigue exhibited high levels of measurement precision and efficiency. J Clin Epidemiol 2013;66.
[13] Fries JF, Krishnan E, Rose M, Lingala B, Bruce B. Improved responsiveness and reduced sample size requirements of PROMIS physical function scales with item response theory. Arthritis Res Ther 2011;13:R147.
[14] Petersen MA, Groenvold M, Aaronson NK, Chie W-C, Conroy T, Costantini A, et al. Development of computerised adaptive testing (CAT) for the EORTC QLQ-C30 dimensions–general approach and initial results for physical functioning. Eur J Cancer 2010;46:1352–8.
[15] Petersen MA, Groenvold M, Aaronson NK, Chie W-C, Conroy T, Costantini A, et al. Development of computerized adaptive testing (CAT) for the EORTC QLQ-C30 physical functioning dimension. Qual Life Res 2011;20:479–90.
[16] Gamper EM, Petersen MA, Aaronson N, Costantini A, Giesinger JM, Holzner B, et al. Development of an item bank for the EORTC role functioning computer adaptive test (EORTC RF-CAT). Health Qual Life Outcomes 2016;14:72.
[17] Gamper E-M, Groenvold M, Petersen MA, Young T, Costantini A, Aaronson N, et al. The EORTC emotional functioning computerized adaptive test: phases I–III of a cross-cultural item bank development. Psychooncology 2014;23:397–403.
[18] Giesinger JM, Petersen MA, Groenvold M, Aaronson NK, Arraras JI, Conroy T, et al. Cross-cultural development of an item list for computer-adaptive testing of fatigue in oncological patients. Health Qual Life Outcomes 2011;9:19.
[19] Petersen MA, Aaronson NK, Chie WC, Conroy T, Costantini A, Hammerlid E, et al. Development of an item bank for computerized adaptive test (CAT) measurement of pain. Qual Life Res 2016;25:1–11.
[20] Petersen MA, Giesinger JM, Holzner B, Arraras JI, Conroy T, Gamper E-M, et al. Psychometric evaluation of the EORTC computerized adaptive test (CAT) fatigue item pool. Qual Life Res 2013;22:2443–54.
[21] Thamsborg LH, Petersen MA, Aaronson NK, Chie WC, Costantini A, Holzner B, et al. Development of a lack of appetite item bank for computer-adaptive testing (CAT). Support Care Cancer 2015;23:1541–8.
[22] Petersen MA, Aaronson NK, Arraras JI, Chie W-C, Conroy T, Costantini A, et al. The EORTC CAT Core—the computer adaptive version of the EORTC QLQ-C30 questionnaire. Eur J Cancer 2018;100:8–16.
[23] Scott NW, Fayers PM, Aaronson NK, Bottomley A, de Graeff A, Groenvold M, et al. A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. J Clin Epidemiol 2009;62:288–95.
[24] Holland PW, Wainer H. Differential item functioning. Routledge; 2012.
[25] Choi SW, Gibbons LE, Crane PK. Lordif: an R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. J Stat Software 2011;39:1.
[26] Jodoin MG, Gierl MJ. Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. Appl Meas Educ 2001;14:329–49.
[27] Cohen J. Statistical power analysis for the behavioral sciences. 1988.
[28] United Nations. World population prospects: the 2017 revision. 2017.
[29] R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012, ISBN 3-900051-07-0. 2014.
[30] IMB Corp. IBM SPSS statistics for windows, version 22.0. Armonk, NY: IBM Corp; 2013.
[31] Hinz A, Singer S, Brahler E. European reference values for the quality of life questionnaire EORTC QLQ-C30: results of a German investigation and a summarizing analysis of six European general population normative studies. Acta Oncol 2014;53:958–65.
[32] Jensen RE, Potosky AL, Moinpour CM, Lobo T, Cella D, Hahn EA, et al. United States population-based estimates of patient-reported outcomes measurement information system symptom and functional status reference values for individuals with cancer. J Clin Oncol 2017;35:1913–20.
[33] Fischer HF, Wahl I, Nolte S, Liegl G, Brahler E, Lowe B, et al. Language-related differential item functioning between English and German PROMIS depression items is negligible. Int J Methods Psychiatr Res 2017;26.
[34] http://www.healthmeasures.net/score-and-interpret/interpret-scores/promis [Accessed 30 August 2018].
[35] Nolte S, Liegl G, Petersen MA, Aaronson NK, Costantini A, Fayers PM, et al. General population normative data for the EORTC QLQ-C30 health-related quality of life questionnaire based on 15,386 persons across 13 European countries, Canada and the Unites States. Eur J Cancer 2018 (in this issue).

2.4 General Population Norm Data for the EORTC QLQ-C30

The primary aim of the international EORTC CAT Core norm data project was to generate European general population norm data to establish the European Norm for the EORTC CAT Core (Liegl et al., 2019). Since the project collected data on all items of the EORTC QLQ-C30, general population norm data for the static cancer core questionnaire could be generated as well. Therefore, new/updated national EORTC QLQ-C30 general population norm data could be generated for 11 EU countries. In addition, national general population norm data could be established for Canada, Russia, Turkey and the U.S., as these countries had been included in the project for comparative purposes.

As described in the Introduction, a major drawback of previous norm data studies had been that the data were only useful for national comparisons, whilst inter-country comparisons were hampered due to the lack of a common sampling methodology across studies. Despite efforts to define a European norm for the EORTC QLQ-C30 by averaging national norm data scores from six European countries (Hinz et al., 2014), the collection of new general population norm data was deemed a priority by the EORTC QLG. The obvious advantage of this new project over previous studies was that a common data collection methodology could be applied across all 15 countries. This common methodology not only allows for intra-country but also for inter-country comparisons and provides current EORTC QLQ-C30 general population norm data for Austria, Canada, Denmark, France, Germany, Hungary, Italy, the Netherlands, Poland, Russia, Spain, Sweden, Turkey, United Kingdom and the U.S. (Nolte, Liegl, et al., 2019).

The following abstract has been taken from the original peer-reviewed article:

Nolte S, Liegl G, Petersen MA, Aaronson NK, Costantini A, Fayers PM, Groenvold M, Holzner B, Johnson CD, Kemmler G, Tomaszewski KA, Waldmann A, Young TE, Rose M. General population normative data for the EORTC QLQ-C30 health-related quality of life questionnaire based on 15,386 persons across 13 European countries, Canada and the Unites States. *European Journal of Cancer*. 2019;107:153–63.
DOI: https://doi.org/10.1016/j.ejca.2018.11.024

> *"OBJECTIVE: The European Organisation for Research and Treatment of Cancer (EORTC) QLQ-C30 health-related quality of life questionnaire is one of the most widely used cancer-specific health-related quality of life questionnaires worldwide. General*

population norm data can facilitate the interpretation of QLQ-C30 data obtained from cancer patients. This study aimed at systematically collecting norm data from the general population to develop European QLQ-C30 norm scores and to generate comparable norm data for individual countries in Europe and North America. METHODS: We collected QLQ-C30 data from the general population across 11 European Union (EU) countries, Russia, Turkey, Canada and United States (n >/= 1000/country). Representative samples were stratified by sex and age groups (18-39, 40-49, 50-59, 60-69 and >/= 70 years). After applying weights based on the United Nations population distribution statistics, we calculated QLQ-C30 domain scores to generate a 'European QLQ-C30 Norm' based on the EU countries. Further, we calculated QLQ-C30 norm scores for all 15 individual countries. RESULTS: A total of 15,386 respondents completed the online survey. For the EU sample, most QLQ-C30 domains showed differences by sex/age, with men scoring somewhat better health than women, while age effects varied across domains. Substantially larger differences were seen in inter-country comparisons, with Austrian and Dutch respondents reporting consistently better health compared with British and Polish respondents. CONCLUSIONS: This study is the first to systematically collect EORTC QLQ-C30 general population norm data across Europe and North America applying a consistent data collection method across 15 countries. These new norm data facilitate valid intra-country as well as inter-country comparisons and QLQ-C30 score interpretation."

European Journal of Cancer 107 (2019) 153–163

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Original Research

General population normative data for the EORTC QLQ-C30 health-related quality of life questionnaire based on 15,386 persons across 13 European countries, Canada and the Unites States

S. Nolte [a,b,*], G. Liegl [a], M.A. Petersen [c], N.K. Aaronson [d], A. Costantini [e], P.M. Fayers [f], M. Groenvold [c,g], B. Holzner [h], C.D. Johnson [i], G. Kemmler [h], K.A. Tomaszewski [j], A. Waldmann [k,l], T.E. Young [m], M. Rose [a,n] on behalf of the EORTC Quality of Life Group

[a] *Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charité - Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany*
[b] *Population Health Strategic Research Centre, School of Health and Social Development, Deakin University, Burwood, VIC, Australia*
[c] *Department of Palliative Medicine, Bispebjerg Hospital, Copenhagen, Denmark*
[d] *Division of Psychosocial Research & Epidemiology, The Netherlands Cancer Institute, Amsterdam, the Netherlands*
[e] *Psycho-Oncology Unit, Sant'Andrea Hospital Sapienza, University of Rome, Rome, Italy*
[f] *Institute of Applied Health Sciences, University of Aberdeen, Aberdeen, United Kingdom*
[g] *Department of Public Health, University of Copenhagen, Copenhagen, Denmark*
[h] *Department of Psychiatry, Psychotherapy and Psychosomatics, Innsbruck Medical University, Innsbruck, Austria*
[i] *University of Southampton, Southampton, United Kingdom*
[j] *Health Outcomes Research Unit, Department of Gerontology, Geriatrics, and Social Work, Faculty of Education, Ignatianum Academy, Krakow, Poland*
[k] *Institute of Social Medicine and Epidemiology, University of Luebeck, Luebeck, Germany*
[l] *Ministry for Health and Consumer Protection, Hamburg Cancer Registry, Hamburg, Germany*
[m] *East & North Hertfordshire NHS Trust Including Mount Vernon Cancer Centre, Northwood, Middlesex, United Kingdom*
[n] *Quantitative Health Sciences, Outcomes Measurement Science, University of Massachusetts Medical School, Worcester, MA, USA*

Received 7 November 2018; accepted 10 November 2018
Available online 19 December 2018

DOI of original article: https://doi.org/10.1016/j.ejca.2018.11.023.
* *Corresponding author*: Sandra Nolte, PhD, Department of Psychosomatic Medicine, Center for Internal Medicine and Dermatology, Charité - Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany.
E-mail address: sandra.nolte@charite.de (S. Nolte).

https://doi.org/10.1016/j.ejca.2018.11.024

KEYWORDS
Quality of life;
EORTC QLQ-C30;
self-report;
patient-reported outcomes;
General population;
norm data;
normative data;
survey;
Europe;
Canada;
USA

Abstract ***Objective:*** The European Organisation for Research and Treatment of Cancer (EORTC) QLQ-C30 health-related quality of life questionnaire is one of the most widely used cancer-specific health-related quality of life questionnaires worldwide. General population norm data can facilitate the interpretation of QLQ-C30 data obtained from cancer patients. This study aimed at systematically collecting norm data from the general population to develop European QLQ-C30 norm scores and to generate comparable norm data for individual countries in Europe and North America.
Methods: We collected QLQ-C30 data from the general population across 11 European Union (EU) countries, Russia, Turkey, Canada and United States ($n \geq 1000$/country). Representative samples were stratified by sex and age groups (18–39, 40–49, 50–59, 60–69 and ≥ 70 years). After applying weights based on the United Nations population distribution statistics, we calculated QLQ-C30 domain scores to generate a 'European QLQ-C30 Norm' based on the EU countries. Further, we calculated QLQ-C30 norm scores for all 15 individual countries.
Results: A total of 15,386 respondents completed the online survey. For the EU sample, most QLQ-C30 domains showed differences by sex/age, with men scoring somewhat better health than women, while age effects varied across domains. Substantially larger differences were seen in inter-country comparisons, with Austrian and Dutch respondents reporting consistently better health compared with British and Polish respondents.
Conclusions: This study is the first to systematically collect EORTC QLQ-C30 general population norm data across Europe and North America applying a consistent data collection method across 15 countries. These new norm data facilitate valid intra-country as well as inter-country comparisons and QLQ-C30 score interpretation.

1. Introduction

The importance of including the patients' voice in clinical practice and research has been recognised widely for several decades [1,2]. In more recent years, patient-reported outcomes (PROs) have also been increasingly acknowledged in drug development and regulatory decision-making [3–5]. Given this growing relevance of PROs, it is not surprising that efforts are undertaken to standardise PRO data, with several initiatives worldwide tackling the issue in different ways. One possible approach—as taken by the International Consortium for Health Outcomes Measurement (ICHOM)—is to define standard sets of existing PRO instruments to enable comparison of outcomes across health-care providers and geographies [6]. A more elaborate approach is to measure PROs by applying modern test theory methods where items measuring the same latent construct are calibrated on the same metric. This serves as the foundation of the application of computer-adaptive tests (CATs) [7–9].

The European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Group (QLG) has a long track record of developing and validating PRO instruments for use in oncology. Their quality of life (QoL) core questionnaire, the QLQ-C30, was developed more than 25 years ago and is one of the most widely used cancer-specific PRO instruments [10,11]. Regarding standardisation, use of the QLQ-C30 as part of clinical routine is frequently recommended in ICHOM standard sets for malignant neoplasms (e.g. [12,13]). Further, the EORTC QLG developed a QLQ-C30 CAT version, the EORTC CAT Core [8]. Through tailoring items to the individual respondent, CATs can achieve the same measurement precision as static instruments while using fewer items; CATs also minimise floor/ceiling effects [14].

While these initiatives are crucial steps toward improving quality and comparability of PRO data, data analysis and interpretation is further enhanced by using a sensible reference for comparative purposes. Depending on the objective of such comparison, a useful reference can be data obtained from cancer patients or norm data collected from the general population. If these data are not available, interpretation of PRO data may be arbitrary. For the comparison of QLQ-C30 scores with cancer-patient data, the EORTC QLG has published reference values generated for various cancer populations [15]. In addition, general population norm data have been collected in different countries over the last two decades, with European data available for Denmark [16,17], Germany [18–20], the Netherlands [21,22], Norway [23,24], Slovenia [25] and Sweden [26,27], with the latest German publication providing a European norm by collating data from different samples [20]. However, a major drawback of this work is that inter-country comparisons are hampered because of the lack of a common sampling methodology across studies.

Given the dearth of research regarding European QLQ-C30 general population norm data for use in inter-country comparisons, the EORTC QLG set out to systematically collect general population data in 13 European countries, Canada and the United States using a common methodology to generate norm data for the QLQ-C30 and the CAT Core [8]. This article presents European QLQ-C30 general population norm data and individual country norms for 15 countries.

2. Material and methods

2.1. Country selection

Country selection was based on several criteria, including population size and balance of geographical location, whilst considering budgetary constraints. The final selection included 11 European Union (EU) countries (Austria, Denmark, France, Germany, Hungary, Italy, the Netherlands, Poland, Spain, Sweden and United Kingdom). For comparative purposes, we also collected data in Russia, Turkey, Canada and the United States.

2.2. Item selection and sociodemographic data

The QLQ-C30 consists of 30 items covering five function subscales (physical, role, emotional, cognitive and social), nine symptom subscales/items (fatigue, nausea/vomiting, pain, dyspnoea, insomnia, appetite loss, constipation, diarrhoea and financial difficulties) and a global health/QoL subscale. Further details of the full-scale survey are reported elsewhere [41]. A range of sociodemographic data were collected, including sex, age, education, employment, relationship status and presence of health conditions.

2.3. Translation and cultural adaptation of included variables

All QLQ-C30 items were available for the languages spoken in the selected countries. In contrast, the sociodemographic variables had to be developed and translated. During this process, assessment of educational attainment proved to be challenging. After a comprehensive consensus process, including review of the International Standard Classification of Education (2011), the final categorisation was deemed adequate by the study collaborators.

2.4. Sampling

To generate sufficiently large sample sizes for estimating stable norms for a range of subgroups, we stratified samples by sex and age group (18–39, 40–49, 50–59, 60–69 and ≥ 70 years) with sample sizes of n = 100/stratum, i.e. n = 1000/country. Sampling of an equal distribution of sex and age groups provides norm data for purposes of comparing cancer-patient data with sex- and age-matched peers from the general population. Further details regarding the rationale for the sample size are reported elsewhere [41].

2.5. Data collection

To ensure a consistent data collection method and representative samples throughout, we subcontracted data collection to GfK SE (www.gfk.com), a panel research company experienced in multinational/multilanguage online surveys. These internet panels are representative for the general population (with internet access) in a given country in terms of sex, age, region, hometown size, household size and socioeconomic status. While GfK achieved most quotas via internet panels, n = 290 respondents (≥ 60 years) provided data via computer-assisted telephone interviews to achieve quotas in Turkey.

Data were collected in March/April 2017. As GfK panel members are registered voluntarily and generally participate when contacted, GfK attains response rates between 75 and 90%.

2.6. Establishing European norm data for the EORTC QLQ-C30

For the final definition of the 'European QLQ-C30 Norm', we included the 11 EU countries. We weighted data by respective country's population size, sex and age distribution based on official 2015 population distribution statistics published by the United Nations [28].

2.7. Statistical analyses

We calculated mean scores and 95% confidence intervals (CIs) of the QLQ-C30 subscales ranging between 0 and 100 following the QLQ-C30 scoring manual [29]. We calculated norm scores for the sex/age strata of the combined EU sample weighted by country size and sex/age distribution. Given the large samples, we further divided the youngest age group into 18–29 and 30–39 years, respectively. Finally, we calculated national norm data for all 15 countries, weighted by respective country's sex/age distribution. For all analyses, we used IBM SPSS Statistics®, version 25.

3. Results

Stratified quotas were achieved in all 15 countries, with n ≥ 100 individuals completing the survey in each sex*age*country subgroup. Country sample sizes ranged between n = 1000 (the Netherlands) and n = 1165 (Spain), leading to a final sample size of N = 15,386 for the full and n = 11,343 for the EU sample (Table 1).

Table 1
Sociodemographic data of full sample (13 European countries, Canada, USA) and "EORTC QLQ-C30 Norm" sample (11 EU countries).

Sociodemographic variable	Full sample (15 countries) N = 15,386		EORTC QLQ-C30 Norm (11 EU countries) n = 11,343	
	n	%	n	%
Sex				
Female	7650	49.7	5623	49.6
Male	7736	50.3	5720	50.4
Age categories (years)				
18–29	1177	7.6	883	7.8
30–39	1902	12.4	1370	12.1
40–49	3049	19.8	2248	19.8
50–59	3059	19.9	2253	19.9
60–69	3138	20.4	2337	20.6
≥70	3061	19.9	2252	19.9
Education				
Less than compulsory education	183	1.2	95	0.9
Compulsory (left school at the minimum school leaving age)	1509	9.9	897	8.0
Some post compulsory (some school after reaching school leaving age without reaching university entrance qualifications (e.g., A-levels))	2050	13.5	1954	17.5
Post compulsory below university (e.g. reaching A levels)	4405	29.0	3408	30.5
University degree (bachelor's degree or equivalent level)	3716	24.4	2689	24.1
Postgraduate degree (master's degree, doctorate or equivalent level)	3337	22.0	2131	19.1
Prefer not to answer[a]	186	–	169	–
Employment status				
Employed full-time	5532	36.2	4087	36.3
Employed part-time	1256	8.2	984	8.7
Homemaker	795	5.2	485	4.3
Student	389	2.5	328	2.9
Unemployed	811	5.3	614	5.5
Retired	5238	34.3	3827	34.0
Self-employed	833	5.5	620	5.5
Other	422	2.8	305	2.7
Prefer not to answer	110	–	93	–
Relationship status				
Single/not in a steady relationship	2589	17.0	1951	17.4
Married or in a steady relationship	10,263	67.4	7640	68.1
Separated/divorced/widowed	2376	15.6	1633	14.5
Prefer not to answer	158	–	119	–
Health status[b,c]				
No health condition/disease	5361	36.6	4204	39.0
Chronic pain	3582	24.5	2468	22.9
Heart disease	1226	8.4	804	7.5
Cancer (excluding basal cell carcinoma)	416	2.8	308	2.9
Depression	1452	9.9	903	8.4
Chronic obstructive pulmonary disease (COPD)	532	3.6	373	3.5
Arthritis	2114	14.4	1427	13.2
Diabetes	1546	10.6	1095	10.2
Asthma	924	6.3	658	6.1
Anxiety disorder	1218	8.3	785	7.3
Obesity	1513	10.3	1072	9.9
Drug/alcohol use disorder	153	1.0	96	0.9
Other	2634	18.0	1863	17.3
Prefer not to answer	631	–	486	–
Country				
Austria[d]	1002	6.5	1002	8.8
Canada	1004	6.5	–	–
Denmark[d]	1003	6.5	1003	8.8
France[d]	1001	6.5	1001	8.8
Germany[d]	1006	6.5	1006	8.9
Hungary[d]	1053	6.8	1053	9.3

Table 1 (*continued*)

Sociodemographic variable	Full sample (15 countries) N = 15,386		EORTC QLQ-C30 Norm (11 EU countries) n = 11,343	
	n	%	n	%
Italy[d]	1036	6.7	1036	9.1
The Netherlands[d]	1000	6.5	1000	8.8
Poland[d]	1024	6.7	1024	9.0
Russia	1007	6.5	–	–
Spain[d]	1165	7.6	1165	10.3
Sweden[d]	1027	6.7	1027	9.1
Turkey	1023	6.6	–	–
United Kingdom[d]	1026	6.7	1026	9.0
USA	1009	6.6	–	–

[a] For the calculation of percentage distributions, the category "prefer not to answer" is treated as missing data.
[b] The sample sizes were reduced by n = 114 (0.7%) in the full sample and n = 79 (0.7%) in the EU subsample, respectively, as respondents had provided implausible data to the question on presence of disease.
[c] The sum of health conditions is larger than the total sample of N = 15,386 (full sample) and n = 11,343 (EU subsample), respectively, as respondents were able to check multiple response options.
[d] Countries included in the calculation of the "EORTC QLQ-C30 Norm".

Sociodemographic characteristics of the two samples were comparable. As per sampling, there was an equal distribution of females/males and age groups. Respondents' age ranged between 18 and 99 years, with mean age 53.6 years. Around 90% of respondents indicated to have at least some post-compulsory education. Across samples, 36% of respondents were working full-time; 34% were retired. About two-thirds reported being married/in a steady relationship. Finally, the most frequently reported diseases were chronic pain, arthritis, diabetes, obesity and depression, with 63% (full sample) and 61% (EU sample), respectively, reporting to have at least one health condition (Table 1).

As shown in Table 2, self-rated function in our EU sample was relatively high. Across subscales, sample mean scores ranged between 84.3 and 86.2 (on a 100-point scale), with 95% CIs between ±0.50 and ± 0.65. The only exception was emotional function, with a mean score of 74.2 (95% CI, ±0.66). Women and men rated themselves similarly, except for emotional function where men rated themselves 4.7 points higher than women. Age effects varied. For physical and role function, women reported decreasing function with increasing age; men did not show age differences. For the remaining function subscales, age effects tended to be in the opposite direction, with older respondents reporting higher function than younger respondents. For some symptom subscales, marked floor effects were observed, with > 80% selecting the lowest/best score for nausea/vomiting, appetite loss, diarrhoea and financial difficulties. Scores ranged between 5.9 (nausea/vomiting) and 29.5 (fatigue), with 95% CIs between ±0.48 and ± 0.79. Men tended to rate themselves lower/better than women, with largest differences observed for insomnia, fatigue and pain. Age effects varied. For pain and dyspnoea, women reported more symptoms with increasing age. In contrast, for fatigue, nausea/vomiting and appetite loss, older respondents tended to score lower/better than younger respondents; for diarrhoea and financial difficulties, this age effect was only seen in men. For global health/QoL, men reported higher scores than women (68.0 for men, 95% CI, ±0.81; 64.3 for women, 95% CI, ±0.74). Respondents aged 50–59 years reported lowest global health/QoL scores (65.8 for men; 62.6 for women; 95% CI, ±1.63 each).

Compared with differences between sex/age groups, inter-country comparisons suggest larger group differences (Table 3, Fig. 1 and 2). Austrian and Dutch respondents reported the best scores, i.e. highest for function and lowest for symptoms. In contrast, Polish and British respondents for the EU sample and Russian, Turkish and United States respondents for the full sample regularly reported worse scores, with differences reaching or exceeding 10 points, a difference that is often applied to indicate clinical relevance [30]. On the global health/QoL subscale, differences between lowest (Poland, Russia, Turkey and United Kingdom) and highest scoring nations (Austria and Netherlands) again exceeded 10 points.

4. Discussion

This study is the first to systematically collect European and individual country general population norm data for the EORTC QLQ-C30 using consistent data collection methods across 15 countries in Europe and North America. The 'European QLQ-C30 Norm' enables valid inter-country comparisons for cancer-patient data. Data from cancer patients can be compared with sex-/age-matched peers from the general population. In addition, the country-specific norm data for 15 countries, especially for those where no QLQ-C30 norm data yet existed, can be used for country-level comparisons.

Table 2
European EORTC QLQ-C30 general population norm data[a]. Mean scores (M)/standard deviations (SD) by subscales stratified by sex and age weighted by sex, age and country according to the United Nations (UN), Department of Economic and Social Affairs population distribution statistics for the year 2015[2].

Domain		Total	Female							Male						
			All female	18-29 years	30-39 years	40-49 years	50-59 years	60-69 years	≥70 years	All male	18-29 years	30-39 years	40-49 years	50-59 years	60-69 years	≥70 years
Function subscales																
Physical function																
	M	85.1	84.3	88.9	86.7	85.8	83.4	82.1	78.5	86.0	85.6	87.3	87.9	86.8	84.9	82.7
	SD	18.9	18.5	14.5	18.0	18.8	18.8	18.7	19.8	19.3	21.6	19.0	18.0	18.2	18.3	19.6
Role function																
	M	84.3	84.1	89.1	84.6	84.1	82.3	83.5	80.7	84.5	82.5	85.2	85.3	84.3	85.4	84.8
	SD	24.6	24.6	20.2	24.6	25.1	25.5	25.3	26.4	24.5	26.3	23.4	24.5	25.1	23.5	22.7
Emotional function																
	M	74.2	71.9	66.2	67.8	69.1	71.0	77.8	79.9	76.6	73.7	71.1	74.3	75.9	82.2	85.7
	SD	24.7	25.3	28.2	26.8	26.4	24.1	21.9	19.8	23.8	26.4	26.5	24.0	23.5	18.9	15.5
Cognitive function																
	M	84.8	84.3	82.8	82.9	82.7	83.2	87.9	86.6	85.2	81.3	84.5	85.7	86.4	87.9	87.7
	SD	21.3	20.9	22.4	23.5	22.7	21.1	16.6	17.2	21.7	27.5	23.3	20.7	20.8	16.6	14.9
Social function																
	M	86.2	85.7	86.1	83.7	83.2	83.8	88.1	89.0	86.7	84.4	84.8	85.3	87.6	89.8	90.2
	SD	24.1	24.6	24.7	26.4	26.8	25.7	22.7	20.4	23.6	26.6	25.2	24.1	22.1	20.8	19.3
Symptom subscales/ items																
Fatigue																
	M	29.5	31.7	34.4	34.6	33.9	32.1	26.6	28.1	27.1	30.7	29.6	27.5	26.7	23.5	21.9
	SD	25.5	25.9	25.3	27.3	27.1	26.2	24.5	24.2	24.8	25.2	25.0	24.5	25.1	23.9	23.3
Nausea/vomiting																
	M	5.9	5.7	7.2	8.5	6.3	4.9	3.7	3.3	6.1	11.9	9.4	5.2	3.4	2.3	1.2
	SD	16.0	14.9	17.4	17.8	16.1	12.8	11.6	11.6	17.1	24.9	20.3	15.0	11.0	9.4	5.5
Pain																
	M	23.5	25.3	20.6	23.3	25.2	28.7	25.4	28.8	21.6	21.3	22.1	21.4	22.9	22.1	19.7
	SD	27.1	27.9	24.9	27.7	28.1	29.2	28.6	28.2	26.0	26.6	25.9	26.0	26.3	26.0	25.0
Dyspnoea																
	M	15.9	16.3	12.6	16.1	16.2	17.0	16.6	19.3	15.5	16.2	15.1	14.0	14.6	16.9	16.4
	SD	24.6	24.5	20.8	23.6	24.8	24.7	25.0	27.3	24.7	26.1	24.7	23.0	23.6	25.5	25.0
Insomnia																
	M	26.6	29.3	26.0	28.9	30.4	35.2	29.2	27.1	23.6	20.4	27.1	26.0	25.6	22.3	20.3
	SD	30.3	30.7	29.8	30.9	31.2	32.5	30.7	28.7	29.6	29.5	30.8	29.8	30.4	28.1	27.5
Appetite loss																
	M	10.0	10.3	11.6	13.5	11.4	9.6	7.6	8.3	9.6	15.8	12.1	7.6	7.5	6.4	5.2
	SD	21.6	21.6	23.2	24.8	23.1	19.7	18.7	18.7	21.6	28.4	22.6	18.1	19.2	17.2	15.8
Constipation																
	M	12.5	14.1	14.1	17.8	14.7	14.1	11.8	12.3	10.9	13.6	12.9	10.0	8.6	9.0	10.1
	SD	23.3	24.4	25.3	27.0	25.3	24.4	21.4	22.3	21.9	25.9	24.1	20.8	19.1	18.9	19.4
Diarrhoea																
	M	9.5	9.0	9.0	12.5	8.8	9.4	6.5	7.7	10.0	14.3	13.5	10.3	8.7	5.9	4.4
	SD	20.9	20.3	20.2	23.5	20.4	20.4	17.0	19.5	21.4	26.4	23.8	21.2	20.2	15.4	13.1
Financial difficulties																
	M	10.6	10.9	9.5	12.9	13.6	12.5	9.3	8.0	10.4	13.6	12.8	11.0	8.9	8.0	5.7
	SD	23.6	24.2	24.1	25.4	27.3	24.8	22.4	20.1	22.9	25.9	24.9	23.9	21.4	20.2	16.7
Global health / Quality of Life																
	M	66.1	64.3	66.4	63.4	62.9	62.6	65.6	64.8	68.0	71.1	67.4	66.3	65.8	67.0	69.6
	SD	21.7	21.8	20.5	21.7	22.9	22.5	22.3	20.9	21.4	21.7	20.9	21.1	22.7	20.8	20.3

[a] The European norm scores for the EORTC QLQ-C30 are based on 11 EU countries (as listed in Table 1).
[b] United Nations, Department of Economic and Social Affairs, Population Division (2017). World Population Prospects: The 2017 Revision, DVD Edition.

To generate these norm data, we subcontracted data collection via online surveys to a panel research company (GfK). Such internet panels are an efficient and cost-effective method to generate norm data, and there is evidence from a comparable study carried out in the context of the United States Patient-Reported Outcomes Measurement Information System (PROMIS) initiative that data are representative of the general population provided that scores are weighted [31], which is consistent with the methods we applied. As 15 individual samples, however, are more heterogenous compared with aforementioned single-country survey, we

Table 3
Country general population norm data for the EORTC QLQ-C30. Mean scores (M)/standard deviations (SD) by subscales weighted by individual country weights and sex and age distributions according to the UN Department of Economic and Social Affairs population distribution statistics for the year 2015[a].

Domain		AUT[b]	CAN	DNK	FRA	DEU	HUN	ITA	NLD	POL	RUS	ESP	SWE	TUR	GBR	USA
Function subscales																
Physical function																
	M	89.7	85.4	84.2	89.1	82.8	89.1	85.2	90.7	81.3	76.3	86.8	88.9	75.8	81.8	80.8
	SD	13.9	19.3	20.4	15.9	21.2	14.0	17.0	14.9	16.5	16.4	16.8	14.6	16.7	23.5	25.2
Role function																
	M	88.9	83.7	82.4	87.8	80.8	87.5	86.1	89.1	83.4	81.0	86.1	88.0	82.3	80.2	81.7
	SD	20.3	25.6	25.9	22.4	27.2	20.5	22.2	21.5	22.1	21.0	21.5	21.4	22.4	29.1	28.2
Emotional function																
	M	78.1	75.5	79.2	76.7	73.9	72.1	73.5	82.3	68.3	68.1	77.1	76.7	65.8	71.0	73.3
	SD	22.3	23.5	25.1	24.3	24.7	22.9	22.7	21.2	25.0	23.7	22.4	21.7	25.5	28.4	28.0
Cognitive function																
	M	89.1	84.7	83.7	86.7	83.9	87.4	87.0	90.3	81.2	79.5	85.7	87.1	75.5	80.5	80.9
	SD	17.8	20.8	22.6	19.5	22.7	17.7	18.6	17.1	22.0	19.0	19.4	18.6	23.2	25.2	25.6
Social function																
	M	92.2	84.9	86.5	90.5	84.8	90.2	88.1	91.9	80.8	83.3	87.8	91.4	83.1	80.3	81.6
	SD	17.1	26.3	24.2	20.8	25.5	19.4	20.6	19.0	25.4	23.7	22.5	19.1	23.6	29.4	29.4
Symptom subscales/ items																
Fatigue																
	M	24.1	29.1	29.9	27.7	31.5	30.2	28.5	23.7	35.9	41.5	23.9	25.6	39.4	32.2	31.9
	SD	22.7	24.1	26.7	26.2	27.2	22.6	23.9	23.0	22.7	23.9	22.7	22.2	24.0	27.6	27.8
Nausea/vomiting																
	M	2.0	6.7	7.9	4.1	6.0	3.8	6.5	3.5	7.4	7.4	4.9	4.0	11.3	8.1	10.9
	SD	8.3	16.6	18.3	13.5	17.2	11.9	15.9	11.8	17.5	15.2	14.5	11.2	18.9	18.9	22.6
Pain																
	M	20.0	24.4	23.4	19.6	27.6	23.5	20.2	17.7	26.0	27.1	22.7	20.4	24.9	26.7	27.5
	SD	24.3	27.2	26.5	24.7	30.9	23.8	23.9	22.9	23.7	23.9	24.0	25.0	22.9	31.2	30.2
Dyspnoea																
	M	10.9	16.3	13.7	14.4	18.7	9.1	15.7	9.5	13.4	23.1	12.4	28.1	18.2	19.5	19.9
	SD	20.6	24.5	23.5	23.8	27.3	19.1	23.0	19.7	21.4	25.0	20.7	26.8	24.8	27.9	28.5
Insomnia																
	M	20.0	30.8	28.5	25.9	27.6	22.1	22.9	21.3	28.6	31.3	25.2	21.8	31.6	32.6	30.8
	SD	27.8	30.6	31.2	30.6	33.1	27.4	27.1	26.1	28.3	29.7	28.0	27.6	28.5	32.8	33.2
Appetite loss																
	M	4.4	11.3	11.8	8.0	10.1	8.0	8.5	4.9	13.0	13.8	9.5	7.6	19.2	14.2	14.1
	SD	16.1	22.8	24.2	19.7	23.3	18.4	19.0	15.1	23.2	22.7	19.9	17.6	24.6	25.2	25.3
Constipation																
	M	6.2	14.6	10.8	11.1	9.6	10.3	14.2	4.9	18.8	14.7	15.3	6.7	23.2	14.7	18.6
	SD	17.3	25.0	23.0	21.2	22.3	20.7	23.4	13.6	26.2	24.4	24.1	17.0	28.6	26.2	28.6
Diarrhoea																
	M	7.5	11.1	10.7	7.3	10.4	9.6	9.3	6.9	12.0	12.2	7.8	7.9	13.5	11.2	13.7
	SD	18.7	21.1	22.0	18.8	22.7	20.2	19.5	17.8	23.3	21.9	18.1	17.2	22.3	23.0	27.1
Financial difficulties																
	M	5.0	12.7	12.2	6.7	11.3	14.8	9.7	4.9	15.5	20.5	9.5	5.8	17.6	14.5	17.5
	SD	17.6	27.0	26.2	19.3	25.0	26.2	21.6	17.1	24.9	29.0	20.7	18.2	25.8	28.7	30.8
Global health / QoL																
	M	75.6	65.9	67.0	68.2	67.0	66.3	64.9	77.4	60.0	59.7	66.8	69.2	60.7	62.3	63.9
	SD	20.0	20.6	23.4	20.1	21.8	20.4	20.3	19.8	20.6	19.7	21.5	22.1	22.7	23.7	22.9

[a] United Nations, Department of Economic and Social Affairs, Population Division (2017). World Population Prospects: The 2017 Revision, DVD Edition.
[b] Country codes: AUT = Austria, CAN=Canada, DNK = Denmark, FRA=France, DEU = Germany, HUN=Hungary, ITA=Italy, NLD=Netherlands, POL=Poland, RUS = Russia, ESP = Spain, SWE=Sweden, TUR = Turkey, GBR = United Kingdom, USA=United Stated of America.

compared our sample characteristics with official population statistics where available. For example, in our EU sample, 6.8% of respondents indicated that they were unemployed (weighted data, not shown), which matches the official 2017 EU unemployment rate of 6.7% for individuals older than 25 years [32]. Further, 64.4% of respondents reported being married/in a steady relationship, which is slightly higher than the EU

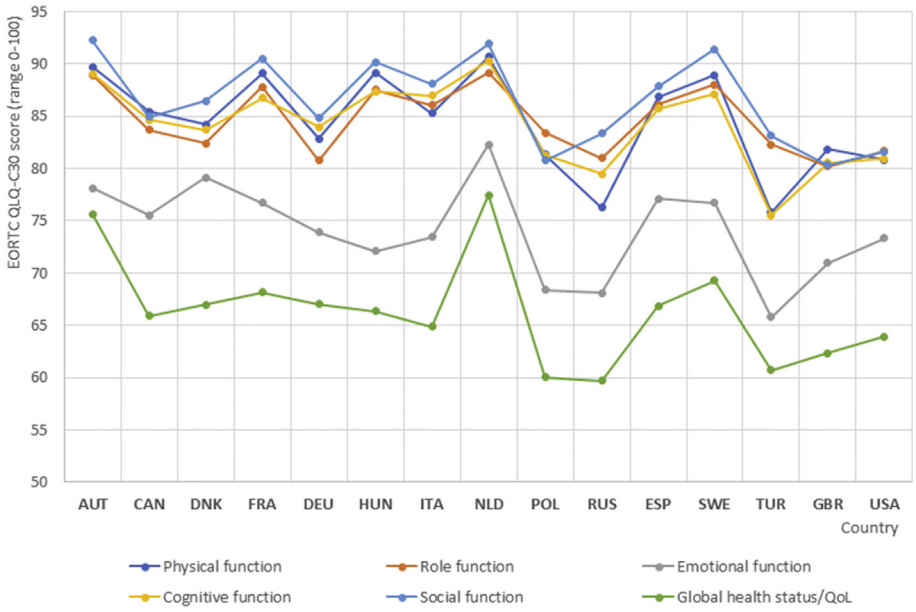

Fig. 1. EORTC QLQ-C30 country reference values for function subscales and global health status/quality of life (for country codes, please refer to Table 3). EORTC QLQ-C30, European Organisation for Research and Treatment of Cancer quality of life core.

average of 59.3% [33]. Finally, self-reported prevalence of several health conditions is largely in-line with prevalence rates published in the literature [34–36]. In contrast, individuals with lower educational levels appear underrepresented in our EU sample, with around 90% reporting at least some post-compulsory education. This is generally lower than in most European countries; however, percentage distribution varies widely by country [37].

Some observed subgroup and country differences warrant further discussion. For example, in emotional function, several symptom subscales and overall QoL, men reported somewhat better scores than women, a finding also observed in other QLQ-C30 norm data studies [20]. Further, in some instances, older respondents reported remarkably high function. For physical function, further subgroup analyses within our oldest age group suggest that decline in self-reported

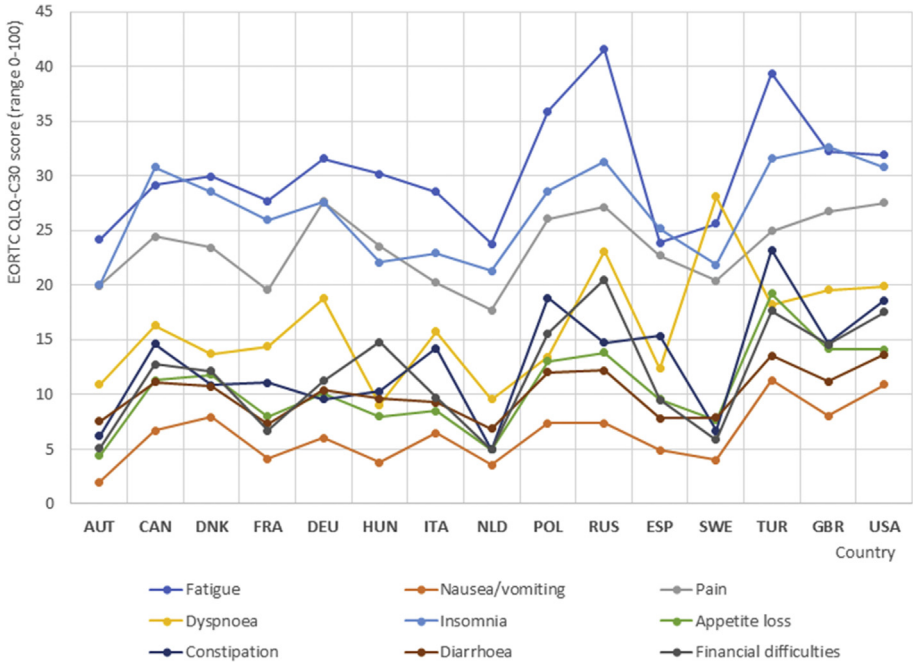

Fig. 2. EORTC QLQ-C30 country reference values for symptom subscales/items (for country codes, please refer to Table 3). EORTC QLQ-C30, European Organisation for Research and Treatment of Cancer quality of life core.

physical function occurs primarily from ≥ 80 years (data not shown). Unfortunately, we did not have sufficient numbers in this age group, as our oldest stratified age group was ≥ 70 years. Also, older respondents' self-reported cognitive function appears high. This finding may be explained by people adjusting health expectations with increasing age. Research also suggests that older persons value different factors compared with younger persons when assessing their health and that younger persons' health perceptions are more affected by health limitations than those of older persons [38]. Additionally, given that in the oldest age group, over 90% were retired, everyday demands on cognitive function may be lower than that of younger respondents, especially those in the workforce. Hence, the construct we are trying to measure may differ depending on respondents' age. The high cognitive function reported by older respondents may also reflect some degree of selection bias, given that respondents had to have internet access and some computer skills. Finally, several reasons might explain the observed country differences. It is conceivable that items and/or response scales have different meanings in different cultures; however, given EORTC's long-standing experience with translations/linguistic validations, it is unlikely that this explains observed differences [39]. It is more likely that differences are indeed true differences between countries and that factors such as the welfare state characteristic [40] play a role in people's self-reported health.

This study has several limitations that should be noted. First, as indicated above, targeting older age groups, i.e. ≥ 80 years, could provide further insight into changes in QoL as a consequence of aging. It was beyond the scope of our study to collect these data as costs are disproportionately high because of the need to often conduct personal/telephone interviews instead of online surveys in older age groups. Second, we observed marked floor effects for several symptoms. Such effects are unavoidable when data are obtained from the general (i.e. relatively healthy) population, especially if fixed-length questionnaires, including some single-item subscales, are used. One possible solution to reduce floor/ceiling effects is the use of CAT such as the EORTC CAT Core [8]. Finally, while our norm data are assumed to be representative of the general population, our sample was relatively highly educated. In our EORTC CAT Core norm data article [41], we explore the influence of educational level on scores. While significant differences were found, with more highly educated respondents reporting better health, the practical relevance of these differences was very small as indicated by small effect sizes. Also, it was difficult to harmonise educational levels across countries; hence, it is plausible that 'post-compulsory' has different meanings in different countries. Despite all of these limitations, online surveys represent an efficient, cost-effective method of obtaining large, representative general population samples. While it comes with disadvantages such as difficulty with reaching older age groups, as seen with the Turkish sample, there are many convincing advantages to this method over personal and telephone interviews (e.g. higher response rates and avoidance of interviewer bias). With the steadily increasing use of the internet in recent years, this method is gaining in popularity [31]. Using internet panels, we were able to obtain a large sample of N = 15,386 persons generating norm data for 15 countries, thereby providing a valuable resource for studies using the QLQ-C30. These general population norm scores are robust, even for stratified analyses, as is evidenced by the generally very small 95% CIs.

5. Conclusions

This study generated European (and North American) and individual country norm data for the EORTC QLQ-C30 based on a common sampling strategy and survey design. We recommend that the 'European QLQ-C30 Norm' be used to compare self-reported health-related quality of life of cancer patients with general population data, especially in multinational projects.

Ethical statement

Ethical approval was not sought as this study is solely based on panel research data. As opposed to medical research where medical professional codes of conduct apply, there is widespread agreement that health research involving volunteers from the general population is not subject to ethical approval. Both the European Pharmaceutical Market Research Association (EphMRA) and the NHS Health Research Authority specify that this type of research does not require ethical approval as long as the research conforms to ethical guidelines. Our online survey was carried out by the panel research company GfK SE which is member of EphMRA. The multinational survey conformed to the required ethical standards by obtaining informed consent from all participants and collecting data completely anonymously. Any identification of the respondents through the authors is impossible.

Acknowledgements

This research was funded by the European Organisation for Research and Treatment of Cancer Quality of Life Group (grant number 001 2015). The work of Bernhard Holzner was partially funded by the Austrian Science Fund (FWF #P26930). The authors thank the many translators who kindly volunteered their time to embark on the task of standardising the sociodemographic variables across languages and cultures. In addition to the collaborators of this project, special thanks go to Susan Bartlett for making the variables in

English and French suitable for the Canadian context; Eveline Bleiker, Jacobien Kieffer, Marieke van Leeuwen for Dutch; Thierry Conroy for French; Agnes Czimbalmos for Hungarian; Alice Iuso for Italian; Tatyana Ionova for Russian; Juan Ignacio Arraras for Spanish; Eva Hammerlid, Yvonne Brandberg for Swedish; Deniz Yüce for Turkish; and Claire Snyder for English for the U.S.-American context.

Conflict of interest statement

None declared

References

[1] Snyder CF, Aaronson NK. Use of patient-reported outcomes in clinical practice. Lancet 2009;374(9687):369–70.
[2] Blazeby JM, Avery K, Sprangers M, Pikhart H, Fayers P, Donovan J. Health-related quality of life measurement in randomized clinical trials in surgical oncology. J Clin Oncol 2006; 24(19):3178–86.
[3] Basch E, Dueck AC. Patient-reported outcome measurement in drug discovery: a tool to improve accuracy and completeness of efficacy and safety data. Expet Opin Drug Discov 2016;11(8): 753–8.
[4] Shields AL, Hao Y, Krohe M, Yaworsky A, Mazar I, Foley C, et al. Patient-reported outcomes in oncology drug labeling in the United States: a framework for navigating early challenges. Am Health Drug Benefits 2016;9(4):188–97.
[5] Kluetz PG, O'Connor DJ, Soltys K. Incorporating the patient experience into regulatory decision making in the USA, Europe, and Canada. Lancet Oncol 2018;19(5):e267–74.
[6] Porter ME, Larsson S, Lee TH. Standardizing patient outcomes measurement. N Engl J Med 2016;374(6):504–6.
[7] Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. Med Care 2007;45(5 Suppl 1): S3–11.
[8] Petersen MA, Aaronson NK, Arraras JI, Chie WC, Conroy T, Costantini A, et al. The EORTC CAT Core-The computer adaptive version of the EORTC QLQ-C30 questionnaire. Eur J Canc 2018;100:8–16.
[9] Fischer HF, Rose M. www.common-metrics.org: a web application to estimate scores from different patient-reported outcome measures on a common scale. BMC Med Res Methodol 2016;16(1):142. https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-016-0241-0.
[10] Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993;85(5):365–76.
[11] Fayers P, Bottomley A. On behalf of the EORTC quality of life group and of the quality of life unit. Quality of life research within the EORTC-the EORTC QLQ-C30. Eur J Canc 2002;38(Suppl 4): S125–33.
[12] Zerillo JA, Schouwenburg MG, van Bommel ACM, Stowell C, Lippa J, Bauer D, et al. An international collaborative standardizing a comprehensive patient-centered outcomes measurement set for colorectal cancer. JAMA Oncol 2017;3(5): 686–94.
[13] Ong WL, Schouwenburg MG, van Bommel ACM, Stowell C, Allison KH, Benn KE, et al. A standard set of value-based patient-centered outcomes for breast cancer: the international Consortium for health outcomes measurement (ICHOM) initiative. JAMA Oncol 2017;3(5):677–85.
[14] Revicki DA, Cella DF. Health status assessment for the twenty-first century: item response theory, item banking and computer adaptive testing. Qual Life Res 1997;6(6):595–600.
[15] Scott NW, Fayers PM, Aaronson NK, Bottomley A, de Graeff A, Groenvold M, et al. EORTC QLQ-C30 reference values. 2008. Brussels, Belgium.
[16] Klee M, Groenvold M, Machin D. Quality of life of Danish women: population-based norms of the EORTC QLQ-C30. Qual Life Res 1997;6(1):27–34.
[17] Juul T, Petersen MA, Holzner B, Laurberg S, Christensen P, Gronvold M. Danish population-based reference data for the EORTC QLQ-C30: associations with gender, age and morbidity. Qual Life Res 2014;23(8):2183–93.
[18] Schwarz R, Hinz A. Reference data for the quality of life questionnaire EORTC QLQ-C30 in the general German population. Eur J Canc 2001;37(11):1345–51.
[19] Waldmann A, Schubert D, Katalinic A. Normative data of the EORTC QLQ-C30 for the German population: a population-based survey. PLoS One 2013;8(9). e74149.
[20] Hinz A, Singer S, Brahler E. European reference values for the quality of life questionnaire EORTC QLQ-C30: results of a German investigation and a summarizing analysis of six European general population normative studies. Acta Oncol 2014; 53(7):958–65.
[21] van de Poll-Franse LV, Mols F, Gundy CM, Creutzberg CL, Nout RA, Verdonck-de Leeuw IM, et al. Normative data for the EORTC QLQ-C30 and EORTC-sexuality items in the general Dutch population. Eur J Canc 2011;47(5):667–75.
[22] Mols F, Husson O, Oudejans M, Vlooswijk C, Horevoorts N, van de Poll-Franse LV. Reference data of the EORTC QLQ-C30 questionnaire: five consecutive annual assessments of approximately 2000 representative Dutch men and women. Acta Oncol 2018:1–11.
[23] Hjermstad MJ, Fayers PM, Bjordal K, Kaasa S. Health-related quality of life in the general Norwegian population assessed by the european organization for research and treatment of cancer core quality-of-life questionnaire: the QLQ=C30 (+ 3). J Clin Oncol 1998;16(3):1188–96.
[24] Hjermstad MJ, Fayers PM, Bjordal K, Kaasa S. Using reference data on quality of life – the importance of adjusting for age and gender, exemplified by the EORTC QLQ-C30 (+3). Eur J Canc 1998;34(9):1381–9.
[25] Velenik V, Secerov-Ermenc A, But-Hadzic J, Zadnik V. Health-related quality of life assessed by the EORTC QLQ-C30 questionnaire in the general slovenian population. Radiol Oncol 2017; 51(3):342–50.
[26] Michelson H, Bolund C, Nilsson B, Brandberg Y. Health-related quality of life measured by the EORTC QLQ-C30–reference values from a large sample of Swedish population. Acta Oncol 2000;39(4):477–84.
[27] Derogar M, van der Schaaf M, Lagergren P. Reference values for the EORTC QLQ-C30 quality of life questionnaire in a random sample of the Swedish population. Acta Oncol 2012;51(1):10–6.
[28] United Nations Department of Economic and Social Affairs Population Division. World population Prospects: the 2017 revision. DVD Edition 2017.
[29] Fayers P, Aaronson N, Bjordal K, Groenvold M, Curran D, Bottomley A, et al. The EORTC QLQ-C30 scoring manual. 3rd ed. Brussels: European Organisation for Research and Treatment of Cancer; 2001.
[30] Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. J Clin Oncol 1998;16(1):139–44.
[31] Liu H, Cella D, Gershon R, Shen J, Morales LS, Riley W, et al. Representativeness of the patient-reported outcomes measurement information System internet panel. J Clin Epidemiol 2010; 63(11):1169–78.

[32] OECD. Harmonised Unemployment Rates (HURs). OECD - updated: june 2018. 2018.
[33] OECD. SF3.3. Cohabitation rate and prevalence of other forms of partnership. 2016.
[34] Breivik H, Collett B, Ventafridda V, Cohen R, Gallacher D. Survey of chronic pain in Europe: prevalence, impact on daily life, and treatment. Eur J Pain 2006;10(4):287–333.
[35] Gallus S, Lugo A, Murisic B, Bosetti C, Boffetta P, La Vecchia C. Overweight and obesity in 16 European countries. Eur J Nutr 2015;54(5):679–89.
[36] Wittchen HU, Jacobi F, Rehm J, Gustavsson A, Svensson M, Jonsson B, et al. The size and burden of mental disorders and other disorders of the brain in Europe 2010. Eur Neuropsychopharmacol 2011;21(9):655–79.
[37] eurostat. Share of the population by level of educational attainment, by selected age groups and country. 2016. 2017.
[38] Idler E, Cartwright K. What do we rate when we rate our health? Decomposing age-related contributions to self-rated health. J Health Soc Behav 2018;59(1):74–93.
[39] Scott NW, Fayers PM, Bottomley A, Aaronson NK, de Graeff A, Groenvold M, et al. Comparing translations of the EORTC QLQ-C30 using differential item functioning analyses. Qual Life Res 2006;15(6):1103–15. discussion 1117-20.
[40] Bergqvist K, Yngwe MÅ, Lundberg O. Understanding the role of welfare state characteristics for health and inequalities – an analytical review. BMC Publ Health 2013;13:1234. 1234.
[41] Liegl G, Petersen MA, Groenvold M, Aaronson NK, Costantini A, Fayers PM, et al. Establishing the European Norm for the health-related quality of life domains of the computer-adaptive test EORTC CAT Core. Eur J Canc 2018 (in this issue).

2.5 Country-specific EORTC QLQ-C30 General Population Norm Data

As part of the European EORTC general population norm data project, there is a unique opportunity to publish detailed national EORTC QLQ-C30 general population norm data for all 15 countries individually, as was done for Germany as presented here. For Germany, national EORTC QLQ-C30 norm data had already been generated by earlier studies (Hinz et al., 2014; Schwarz & Hinz, 2001; Waldmann et al., 2013). Therefore, the updated norm data lend themselves to a comparison with the results of these earlier publications.

It was found that the general population norm data studies published by Schwarz and Hinz (2001) and Hinz et al. (2014) reported better EORTC QLQ-C30 scores (i.e., higher functioning and lower symptom scores) for all 15 scales compared to the updated German general population norm data (Nolte et al., 2020), with largest differences seen in males. In contrast, the German norm data published by Waldmann et al. (2013) were largely in line with the updated norm data. To explore potential reasons for the observed discrepancies, especially between the updated norm data and those published by Schwarz and Hinz (2001) and Hinz et al. (2014), data from a large German health monitoring survey (Steppuhn et al., 2017; Thom et al., 2017) were examined for comparative purposes. The comparison suggested that the German health monitoring survey showed similar or slightly worse 12-month prevalence data for various chronic conditions compared with the present project thereby providing support for the representativeness of the newly collected data. It was concluded that especially Schwarz and Hinz (2001) and Hinz et al. (2014) may have recruited too many individuals that were in good health, i.e., both samples were likely healthier on average than the German general population. Based on the assumption that the newly collected EORTC QLQ-C30 norm data were obtained from a sample that was more representative of the German general population than those samples contained in previous studies, it was recommended that the new EORTC QLQ-C30 general population norm data be used and replace all previously published norm data for Germany (Nolte et al., 2020).

The following abstract has been taken from the original peer-reviewed article:

Nolte S*, Waldmann A*, Liegl G, Petersen MA, Groenvold M, Rose M. Updated EORTC QLQ-C30 general population norm data for Germany. *European Journal of Cancer*. 2020;137:161–70. (*joint first authorship) DOI: https://doi.org/10.1016/j.ejca.2020.06.002

"OBJECTIVE: The European Organisation for Research and Treatment of Cancer (EORTC) core questionnaire, QLQ-C30, is a frequently used patient-reported outcome (PRO) instrument to assess health-related quality of life of patients with cancer. To enhance the understanding and interpretation of PRO data, it is important to obtain norm data from the general population. This article presents updated general population norm data for the EORTC QLQ-C30 for Germany. METHODS: Data were obtained as part of a larger study collecting EORTC QLQ-C30 norm data across 15 countries via an online survey. After linear transformation of EORTC QLQ-C30 raw scores, data were weighted based on the United Nations' population distribution statistics. Data are presented by age and sex/age. RESULTS: A total of 1006 Germans responded to the survey. Across EORTC QLQ-C30 domains, different response patterns were observed, with men generally scoring better, that is, higher in most function scales and lower in most symptom scales/items than women. For age, mixed patterns were observed. While older respondents scored worse/lower in physical and role functioning, emotional functioning scores appeared to increase with increasing age. For the symptom scales/items, some symptoms were relatively stable across age groups, while others either increased or decreased with increasing age. CONCLUSIONS: This study presents updated EORTC QLQ-C30 general population norm data for Germany that can readily be used for comparative purposes with data obtained from patients with cancer."

European Journal of Cancer 137 (2020) 161–170

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Original Research

Updated EORTC QLQ-C30 general population norm data for Germany

Sandra Nolte [a,b,*,1], Annika Waldmann [c,d,1], Gregor Liegl [a], Morten Aa Petersen [e], Mogens Groenvold [e,f], Matthias Rose [a] on behalf of the EORTC Quality of Life Group

[a] *Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin Institute of Health, Medical Department, Division of Psychosomatic Medicine, Berlin, Germany*
[b] *School of Health and Social Development, Faculty of Health, Deakin University, Burwood, VIC, Australia*
[c] *Institute of Social Medicine and Epidemiology, University of Luebeck, Luebeck, Germany*
[d] *Hamburg Cancer Registry, Hamburg, Germany*
[e] *Department of Palliative Medicine, Bispebjerg Hospital, Copenhagen, Denmark*
[f] *Department of Public Health, University of Copenhagen, Copenhagen, Denmark*

Received 23 May 2020; accepted 1 June 2020
Available online 7 August 2020

KEYWORDS
Quality of life;
EORTC QLQ-C30;
Self-report;
Patient-reported outcomes;
General population;
Norm data;
Normative data;
Survey;
Germany

Abstract ***Objective:*** The European Organisation for Research and Treatment of Cancer (EORTC) core questionnaire, QLQ-C30, is a frequently used patient-reported outcome (PRO) instrument to assess health-related quality of life of patients with cancer. To enhance the understanding and interpretation of PRO data, it is important to obtain norm data from the general population. This article presents updated general population norm data for the EORTC QLQ-C30 for Germany.

Methods: Data were obtained as part of a larger study collecting EORTC QLQ-C30 norm data across 15 countries via an online survey. After linear transformation of EORTC QLQ-C30 raw scores, data were weighted based on the United Nations' population distribution statistics. Data are presented by age and sex/age.

Results: A total of 1006 Germans responded to the survey. Across EORTC QLQ-C30 domains, different response patterns were observed, with men generally scoring better, that is, higher in most function scales and lower in most symptom scales/items than women. For age, mixed patterns were observed. While older respondents scored worse/lower in physical and role functioning, emotional functioning scores appeared to increase with increasing age. For the symptom scales/items, some symptoms were relatively stable across age groups, while

* *Corresponding author:* Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin Institute of Health, Medical Department, Division of Psychosomatic Medicine, Charitéplatz 1, 10117, Berlin, Germany.
E-mail address: sandra.nolte@charite.de (S. Nolte).
[1] shared first authorship.

https://doi.org/10.1016/j.ejca.2020.06.002

others either increased or decreased with increasing age.
Conclusions: This study presents updated EORTC QLQ-C30 general population norm data for Germany that can readily be used for comparative purposes with data obtained from patients with cancer.

1. Introduction

Many short- and long-term constraints and side-effects resulting from cancer and its treatment are related to the subjective experience of the individual. Therefore, the assessment of patient-reported outcome (PRO) data is becoming increasingly important not only in oncological practice and research but also in the development and regulatory evaluation of cancer drugs [1–3]. However, the measurement and interpretation of patients' self-reported health-related quality of life (HRQoL) scores, a subtype of PRO data, come with methodological challenges, such as score interpretation. For example, to interpret scores and to understand how specific scores compare with other groups, a sensible comparator should be selected (e.g. other patients with cancer, norm data obtained from the general population and so on).

The European Organisation for Research and Treatment of Cancer (EORTC) core questionnaire, QLQ-C30, is a frequently used PRO instrument to assess HRQoL of patients with cancer. It covers 15 domains/aspects of HRQoL, including a global health/quality of life (QoL) score [4]. For the understanding and interpretation of EORTC QLQ-C30 data by use of a comparator, several efforts have already been undertaken. For example, for the comparison of EORTC QLQ-C30 data with data from other patients with cancer, reference values from various oncological populations exist [5]. Further, general population norm data are available for several European countries [6–14], including four studies on the German general population [15–18].

Despite the great efforts to generate general population norm data, the publication of various normative samples since the late 1990s has led to several challenges. At the national level, for example, different sampling strategies were applied across the four German studies, leading to differences between the studies [15–18]. In the context of multinational studies, uniform general population norm data are even more difficult to establish, unless HRQoL scores are obtained as part of a concerted data collection effort. A first attempt to provide summary European norm data was undertaken in 2014 by collating data from previously published studies [17]. However, this summary publication was not able to overcome the limitation of differences in sampling methods between the different studies. Consequently, to date, comparability of general population norm data between countries has been limited.

Owing to the lack of high-quality European general population norm data for the EORTC QLQ-C30, the EORTC Quality of Life Group decided to fund a large-scale cross-country project to collect norm data using a common methodology across 13 European countries. The aim of the study was to generate up-to-date norm data for the EORTC QLQ-C30 in a representative sample of the adult population in Europe. In addition, norm data were collected in Canada and USA for comparative purposes [19]. In this article, we describe the general population norm data sample collected from the German adult population. While national-level data have already been published as part of the core article [19], this article provides a more detailed overview of the German general population norm data by stratifying the sample by sex and age/sex. In addition, this article discusses similarities and differences compared with the previously published German general population samples that are adjusted using the same weighting procedure as for our data.

2. Material and methods

2.1. Item selection and socio-demographic data

This study was part of a larger study aimed at generating European general population norm data for the EORTC QLQ-C30 and its computerised adaptive test (CAT) version, the EORTC CAT Core [20]. The main study collected norm data in 15 countries in Europe and North America [19,21].

The 30 items of the EORTC QLQ-C30 cover 15 domains, of which five are function scales, nine are symptom scales/items and one is a global health/QoL scale. In addition, we collected data on sex, age, educational attainment, relationship status, employment status, presence of a range of doctor-diagnosed health conditions/diseases and ethnicity.

2.2. Sampling and data collection

Samples were stratified by sex and age. We defined five age groups (18–39, 40–49, 50–59, 60–69, ≥70 years), leading to 10 strata (female/male*age groups), with an anticipated sample size of n = 100/stratum. Data were collected via internet panels by GfK SE (www.gfk.com), a panel research company with long-standing experience with international surveys. GfK warrants panels to be

Table 1
Socio-demographic data of full German sample and sample stratified by sex (crude sample).

Sociodemographic variable	Full sample (n = 1006)		Females (n = 501)		Males (n = 505)	
	n	%[a]	n	%[a]	n	%[a]
Age (years)						
(mean, SD)	53.8	15.0	53.5	15.2	54.2	14.8
Age (years, in categories)						
18–39	200	19.9	100	20.0	100	19.8
40–49	201	20.0	100	20.0	101	20.0
50–59	201	20.0	101	20.2	100	19.8
60–69	201	20.0	100	20.0	101	20.0
≥70	203	20.2	100	20.0	103	20.4
Education						
Less than compulsory education	1	0.1	1	0.2	0	0.0
Compulsory (left school at the minimum school leaving age)	112	11.2	56	11.2	56	11.2
Some postcompulsory (some school after reaching school leaving age without reaching university entrance qualification)	396	39.7	222	44.5	174	34.9
Postcompulsory below university	163	16.3	85	17.0	78	15.6
University degree (bachelor's degree or equivalent level)	125	12.5	52	10.4	73	14.6
Postgraduate degree (master's degree, doctorate or equivalent level)	201	20.1	83	16.6	118	23.6
Prefer not to answer	8		2		6	
Employment status						
Employed full time	370	37.1	127	25.6	243	48.4
Employed part time	111	11.1	90	18.1	21	4.2
Homemaker	43	4.3	39	7.9	4	0.8
Student	24	2.4	17	3.4	7	1.4
Unemployed	31	3.1	17	3.4	14	2.8
Retired	333	33.4	171	34.5	162	32.3
Self-employed	60	6.0	23	4.6	37	7.4
Other	26	2.6	12	2.4	14	2.8
Prefer not to answer	8		5		3	
Relationship status						
Single/not in a steady relationship	170	17.1	76	15.3	94	18.8
Married or in a steady relationship	652	64.4	296	59.7	356	71.1
Separated/divorced/widowed	175	17.6	124	25.0	51	10.2
Prefer not to answer	9		5		4	
Health status[b]						
No health condition/disease	345	36.9	158	33.6	187	40.3
Chronic pain	270	28.9	156	33.2	114	24.6
Heart disease	74	7.9	32	6.8	42	9.1
Cancer (excluding basal cell carcinoma)	31	3.3	21	4.5	10	2.2
Depression	77	8.2	40	8.5	37	8.0
Chronic obstructive pulmonary disease (COPD)	34	3.6	15	3.2	19	4.1
Arthritis	152	16.3	99	21.1	53	11.4
Diabetes	118	12.6	47	10.0	71	15.3
Asthma	65	7.0	44	9.4	21	4.5
Anxiety disorder	37	4.0	17	3.6	20	4.3
Obesity	83	8.9	47	10.0	36	7.8
Drug/alcohol use disorder	10	1.1	1	0.2	9	1.9
Other	175	18.7	100	21.3	75	16.2
Prefer not to answer	*70*		*30*		*40*	

[a] Percentage excludes those who preferred not to answer respective question.
[b] Sum of health conditions/diseases is larger than the total sample of n = 1006, as respondents were able to check multiple response options.

representative for the general population in a given country. Data collection took place in March/April 2017. Further details on sampling, choice of countries, stratification and so on are reported elsewhere [19].

2.3. Statistical analyses

Socio-demographic data were analysed descriptively. Calculated mean scores of the 15 EORTC QLQ-C30 subscales were transformed to a range between 0 and 100 [4]. As a rough guide to determine group differences, we applied a cut-off of ≥10 points to indicate moderate group differences [22].

As the chosen sampling strategy was based on an equal number of subjects per sex/age stratum (Refer to Sampling and data collection), reported means based on the total sample were weighted by Germany's sex/age distribution to achieve that the 'German norm' was as

representative as possible of the German general population. Weights were derived from population distribution statistics for the year 2015 as published by the United Nations, Department of Economic and Social Affairs, Population Division population distribution statistics [23]. To enable accurate comparison of the new norm data with previously published German norm data [15–17], the weighted sex/age structure of our population was used to adjust the previously reported norm scores following the Hjermstad *et al.* [10] procedure.

We used IBM SPSS Statistics®, version 25, for all analyses.

3. Results

3.1. Sample description

As shown in Table 1 (crude sample), the sample consisted of 501 women and 505 men. The mean age was 54 years. Approximately 11% had less than or compulsory education, while about one third had a university or postgraduate degree. About one third was working full time, and one third was retired. About two thirds were married/in a steady relationship. Sixty-three percent of study participants reported having a doctor-diagnosed health condition/disease, with the most frequently reported diseases being chronic pain (29%), arthritis (16%) and diabetes (13%). As expected for a representative German sample with the given age structure [24], men had a higher educational level, and more men were in full-time employment compared with women, while more women than men worked either part time or were homemakers. Furthermore, more women reported to be separated/divorced/widowed and had at least one health condition compared with men (Refer toTable 1 for crude sample, Supplement Table 1 for weighted sample).

3.2. Overall HRQoL in Germany (weighted, unweighted)

Weighted mean scores for the function scales ranged between 73.9 (emotional functioning) and 84.8 (social functioning), while it was 67.0 for the global health/QoL scale. Symptom scores ranged from 6.0 for nausea/vomiting to 31.5 for fatigue.

To assess the impact of weighting on mean scores, weighted and unweighted scores were compared. As shown in Table 2, respective mean scores did not divert by more than 1.5 points showing minimal impact of the weighting procedure on the norm mean scores.

3.3. HRQoL by age

As shown in Table 3, data stratification by age suggested that physical and role functioning tended to deteriorate with increasing age, while particularly for emotional functioning, the reverse seemed to be the case. In the case of the latter, however, the youngest age group reported a relatively low level of emotional functioning compared with the other four function scales, with more than 10 points difference between respective subscale's mean score. For cognitive functioning, the youngest age group scored at least 5 points lower than any of the other age groups, with highest scores observed in the age group of 60–69 years. For global health/QoL, younger respondents showed highest scores (71.4 points), which monotonously decreased to 63.9 points reported by the oldest age group.

For symptoms, largest age differences were observed for pain, dyspnoea and insomnia, with younger respondents reporting substantially lower scores than older age groups. In contrast, the youngest age group reported higher nausea/vomiting symptom burden; they also tended to show highest symptom burden in appetite loss, constipation, diarrhoea and financial difficulties compared with older respondents with largely monotonous decreases of symptom burden from young to old.

3.4. HRQoL by sex and age

In addition to stratification by age, we further divided the sample into women (Table 4a) and men (Table 4b).

Table 2
EORTC QLQ-C30 general population norm data for Germany. Mean scores (M)/standard deviations (SD) by scales/symptoms, comparison of weighted[a] and unweighted scores.

Domain	Weighted		Unweighted	
	M	SD	M	SD
Function subscales				
Physical functioning	82.8	21.2	82.0	21.5
Role functioning	80.8	27.2	80.3	27.4
Emotional functioning	73.9	24.7	75.1	24.2
Cognitive functioning	83.9	22.7	85.4	21.1
Social functioning	84.8	25.5	85.1	25.5
Symptom subscales/items				
Fatigue	31.5	27.2	31.4	27.7
Nausea/vomiting	6.0	17.2	5.2	15.7
Pain	27.6	30.9	28.3	31.1
Dyspnoea	18.7	27.3	19.6	27.8
Insomnia	27.6	33.1	28.9	33.6
Appetite loss	10.1	23.3	9.3	22.2
Constipation	9.6	22.3	8.9	21.6
Diarrhoea	10.4	22.7	9.7	22.2
Financial difficulties	11.3	25.0	10.4	24.1
Global health/Quality of Life	67.0	21.8	65.9	22.2

EORTC QLQ-C30, European Organisation for Research and Treatment of Cancer core questionnaire.
[a] Weighted by sex and age according to the United Nations, Department of Economic and Social Affairs, population distribution statistics for the year 2015 (United Nations, Department of Economic and Social Affairs, Population Division (2017). World Population Prospects: The 2017 Revision, DVD Edition).

Table 3
EORTC QLQ-C30 general population norm data for adults in Germany. Mean scores (M)/standard deviations (SD) by scales/symptoms stratified by age group (weighted data).

Domain	Total (i.e. men and women combined)											
	Total		18–39 years		40–49 years		50–59 years		60–69 years		≥70 years	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Function subscales												
Physical functioning	82.8	21.2	87.0	19.7	87.7	16.9	82.9	21.3	78.9	21.9	73.6	23.4
Role functioning	80.8	27.2	84.7	25.1	84.6	25.3	80.2	26.5	77.7	28.4	73.0	30.3
Emotional functioning	73.9	24.7	69.4	26.5	73.8	25.1	71.9	25.6	79.1	21.7	80.3	19.7
Cognitive functioning	83.9	22.7	79.8	26.9	86.2	20.4	84.8	21.0	88.9	16.6	84.9	20.5
Social functioning	84.8	25.5	84.5	25.9	84.0	26.9	83.8	25.5	86.7	25.3	85.9	23.7
Symptom subscales/items												
Fatigue	31.5	27.2	31.0	25.3	33.5	27.7	30.8	27.3	29.2	28.1	33.1	29.1
Nausea/vomiting	6.0	17.2	9.7	22.2	4.2	15.0	4.9	13.6	4.0	13.1	3.6	12.6
Pain	27.6	30.9	21.9	29.0	24.4	29.0	29.9	30.3	32.8	32.9	35.0	33.1
Dyspnoea	18.7	27.3	13.7	23.4	16.1	24.5	18.9	27.2	23.2	29.2	27.0	32.2
Insomnia	27.6	33.1	22.1	30.7	26.9	32.1	32.1	36.3	30.2	34.3	32.0	33.0
Appetite loss	10.1	23.3	12.8	26.7	8.8	22.5	8.6	20.4	8.2	21.1	9.4	21.3
Constipation	9.6	22.3	12.3	24.9	8.4	23.1	8.6	19.8	8.2	20.5	7.5	19.8
Diarrhoea	10.4	22.7	12.7	23.7	8.5	21.4	11.8	25.6	7.4	18.8	8.9	21.1
Financial difficulties	11.3	25.0	14.0	28.0	10.8	25.4	9.9	21.9	9.5	24.2	9.3	22.1
Global health/Quality of Life	67.0	21.8	71.4	19.8	66.6	20.9	64.4	22.4	64.3	24.9	63.9	21.7

EORTC QLQ-C30, European Organisation for Research and Treatment of Cancer core questionnaire.

The observed decrease of both self-reported physical and role functioning but increase in emotional functioning with increasing age was observed in both women and men. However, observed low scores for emotional functioning for the youngest age group were particularly pronounced in women who reported >15 points difference between this subscale and any of the remaining function subscales. For cognitive functioning, the observed peak in the age group of 60–69 years was only apparent in women, while the observed lower scores in cognitive functioning reported by the youngest age group were only seen in men. For symptoms, observed age differences regarding dyspnoea and insomnia were only apparent in women. In contrast, the observed age difference regarding higher levels of nausea/vomiting, appetite loss, constipation, diarrhoea and financial difficulty of the youngest age group compared with most other age groups was only found in men.

When comparing self-reported health between women and men globally, men scored slightly higher/better on the global health/QoL scale and the function scales, except for cognitive and social functioning compared with women. Men also scored lower/better on most symptom scales/items, except for nausea/vomiting and diarrhoea. However, when comparing respective total mean scores, none of these differences reached the *a priori* defined 10-point threshold, and only one subscale showed a difference of >5 points (pain, lower for men). When exploring each age stratum, however, some larger group differences were observed. For example, large differences were seen in the youngest age group in cognitive functioning with women scoring almost 10 points higher/better than men. Furthermore, women aged 50–59 years reported substantially lower/worse physical and role functioning and higher/worse symptom burden for insomnia, pain and dyspnoea than men of the same age group. Group differences in the next older age group (60–69 years) were substantially smaller, with the only marked difference observed for insomnia in favour of men.

Stratified results by sex and age are further shown in Fig. 1a/b (women) and Fig. 2a/b (men) for easier visualisation of the findings.

4. Discussion

In this article, we present updated general population norm data for Germany, which were obtained as part of a large, multinational study collecting norm data across 15 countries in Europe and North America, thereby applying a consistent data collection method throughout [19]. The application of a common methodology is particularly crucial to ensure that data can be compared across countries and cultures. For the purpose of this study, data collection was subcontracted to the panel research company GfK. Collecting data via internet panels is an efficient, cost-effective method to generate norm data, and there is evidence from a comparable study carried out by the Patient-Reported Outcomes Measurement Information System initiative that these data are representative of the general population. Nevertheless, this same group recommends weighting of scores to obtain a truly representative sample [25], a method which is consistent with our procedure. While weighting did not have a substantial impact on obtained mean scores in our study, the direction of score

Table 4a
EORTC QLQ-C30 general population norm data for women in Germany. Mean scores (M)/standard deviations (SD) by scales/symptoms stratified by age group (weighted data).

Domain	Women											
	Total		18–39 years		40–49 years		50–59 years		60–69 years		≥70 years	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Function subscales												
Physical functioning	81.4	21.3	88.0	17.4	86.6	18.3	78.0	22.4	78.3	23.0	72.1	22.8
Role functioning	79.7	27.5	86.7	22.5	84.2	24.9	75.4	27.4	75.8	29.9	71.3	31.7
Emotional functioning	73.1	25.3	68.7	26.4	73.8	27.4	70.1	26.2	77.2	23.3	79.3	20.9
Cognitive functioning	85.2	21.2	84.9	22.5	85.5	22.7	82.5	20.1	90.5	14.3	84.3	22.1
Social functioning	85.7	25.6	87.7	23.8	84.7	27.9	81.0	27.5	84.7	28.7	88.0	22.1
Symptom subscales/items												
Fatigue	33.8	27.9	32.5	25.1	36.7	30.3	34.4	28.5	30.2	28.8	35.3	29.0
Nausea/vomiting	5.5	16.2	6.4	17.3	4.2	16.5	5.9	14.8	5.7	16.5	4.7	15.4
Pain	30.7	32.1	21.9	27.7	28.2	31.2	37.6	33.6	33.7	34.3	38.7	33.3
Dyspnoea	19.3	28.5	9.2	19.2	17.0	26.2	23.8	29.6	22.7	29.6	30.7	34.7
Insomnia	29.7	33.5	19.7	27.8	27.7	33.9	39.9	37.1	35.0	35.7	34.3	33.0
Appetite loss	10.7	23.4	10.1	24.4	10.7	25.0	10.6	20.5	10.0	24.0	12.0	23.0
Constipation	9.9	22.9	10.8	23.9	10.3	25.4	9.2	20.6	9.7	22.4	8.7	21.5
Diarrhoea	9.5	22.5	9.8	21.9	5.7	19.0	11.6	25.2	7.0	19.8	12.0	24.8
Financial difficulties	11.6	25.6	11.7	27.6	12.0	27.5	13.2	23.6	11.0	26.5	10.0	22.0
Global health/Quality of Life	65.8	22.0	72.3	18.2	65.3	22.3	60.1	22.8	63.8	26.6	62.6	21.0

EORTC QLQ-C30, European Organisation for Research and Treatment of Cancer core questionnaire.

Table 4b
EORTC QLQ-C30 general population norm data for men in Germany. Mean scores (M)/standard deviations (SD) by scales/symptoms stratified by age group (weighted data).

Domain	Men											
	Total		18–39 years		40–49 years		50–59 years		60–69 years		≥70 years	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Function subscales												
Physical functioning	84.3	21.0	86.0	21.6	88.7	15.4	87.8	19.1	79.5	20.8	75.7	24.2
Role functioning	82.0	26.8	82.7	27.3	85.0	25.9	85.0	24.8	79.7	26.8	75.2	28.3
Emotional functioning	74.7	24.0	70.1	26.6	73.7	22.7	73.7	25.0	81.2	19.8	81.7	18.0
Cognitive functioning	82.6	24.1	75.0	29.8	87.0	17.9	87.0	21.8	87.3	18.8	85.8	18.3
Social functioning	83.9	25.5	81.4	27.5	83.3	26.1	86.5	23.2	88.8	21.3	83.0	25.7
Symptom subscales/items												
Fatigue	29.1	26.2	29.6	25.5	30.5	24.6	27.1	25.8	28.1	27.4	30.0	29.1
Nausea/vomiting	6.6	18.1	12.8	25.8	4.3	13.5	3.8	12.3	2.3	7.9	2.1	7.3
Pain	24.3	29.3	21.9	30.2	20.8	26.4	22.2	24.5	31.8	31.6	29.9	32.3
Dyspnoea	18.1	26.0	18.0	26.2	15.2	22.9	14.0	23.8	23.8	28.9	22.0	27.9
Insomnia	25.4	32.6	24.3	33.2	26.1	30.4	24.3	33.8	25.1	32.2	28.8	33.1
Appetite loss	9.6	23.2	15.4	28.6	6.9	19.6	6.7	20.1	6.3	17.5	5.8	18.4
Constipation	9.2	21.7	13.8	25.8	6.6	20.6	8.0	19.0	6.6	18.3	5.8	17.1
Diarrhoea	11.4	22.8	15.4	25.0	11.2	23.2	12.0	26.2	7.9	17.8	4.5	13.3
Financial difficulties	10.9	24.4	16.2	28.2	9.6	23.3	6.7	19.5	7.9	21.7	8.4	22.3
Global health/Quality of Life	68.2	21.5	70.5	21.3	68.0	19.5	68.7	21.3	64.8	23.2	65.7	22.6

EORTC QLQ-C30, European Organisation for Research and Treatment of Cancer core questionnaire.

adjustments was consistent with the higher weights given to older strata compared with younger strata, hence, consistent with age group differences reported in the Results section of this article.

Some of our findings are in line with response patterns that would be expected, such as the observed decline in physical and role functioning with increasing age. In contrast, comparatively low scores in cognitive functioning in young men were unexpected. However, this finding is in line with other studies suggesting that older adults frequently show a disconnection between subjective and objective memory performance and subsequently overestimate their cognitive functioning [26]. Coupled with possible downward comparison of the older ages groups [27], that is, comparison of oneself to people of the same age who are less cognitively able, as

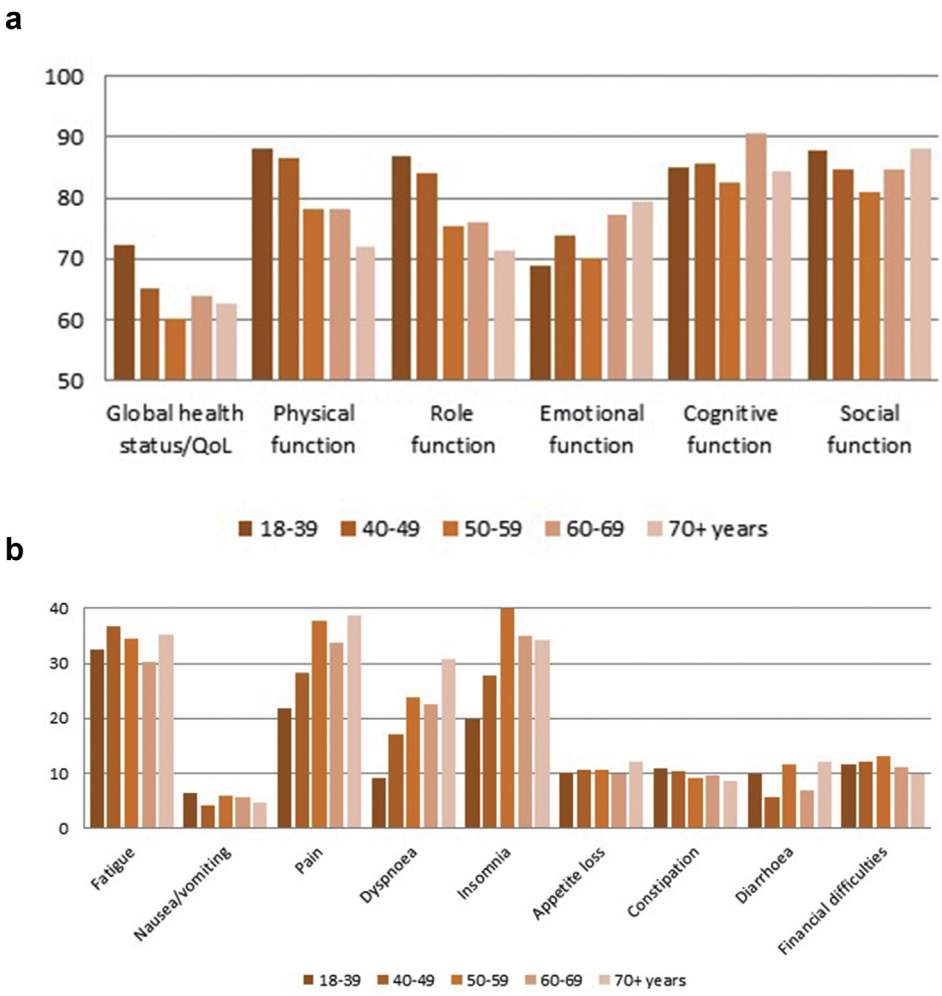

Fig. 1. a) German EORTC QLQ-C30 general population normative data for global health status/quality of life and function subscales for women by age group. b) German EORTC QLQ-C30 general population normative data for symptom subscales and items for women by age group. EORTC QLQ-C30, European Organisation for Research and Treatment of Cancer core questionnaire.

well as with lower demands due to retirement, this might explain the relatively high self-reported cognitive functioning in older respondents. As our data were obtained via an online survey, it is also possible that older participants were biased towards those who were sufficiently (cognitively) capable and healthy to respond to an electronic survey.

For the EORTC QLQ-C30 symptom scales, some response patterns were again in the expected direction (e.g. pain and dyspnoea), while other symptom scales did not show any obvious trend. One exception was the observation of some consistently higher symptom burden in younger men (i.e. nausea/vomiting, appetite loss, constipation and diarrhoea) which may be related to differences in leisure activities (e.g. the high incidence of binge drinking, especially in younger men, might explain some of the findings [28]). In the context of symptom scales, however, it needs to be stressed that our sample is based on respondents from the general population and many symptoms are included in the EORTC QLQ-C30 because of particular relevance to patients with cancer during or after treatment. Therefore, some floor effects in the symptom scales in particular – regardless of age – can be expected when seeking responses to these items from the general population.

As earlier publications already established general population norm data for Germany, we compared our data to these publications [15–17]. Of note, we applied the same weights to the results from the earlier publications as applied to the data in this article to ensure comparability. First, it is striking that the general population data published by Schwarz and Hinz [15] and Hinz *et al.* [17] reported better scores (i.e. higher functioning/lower symptom scores) for all 15 EORTC QLQ-C30 scales than found in our data. For women, one function and four symptom scales/items showed deviations of >10 points. Even greater differences were seen for men, with four function and three symptom scales/items showing a deviation of >10 points. In

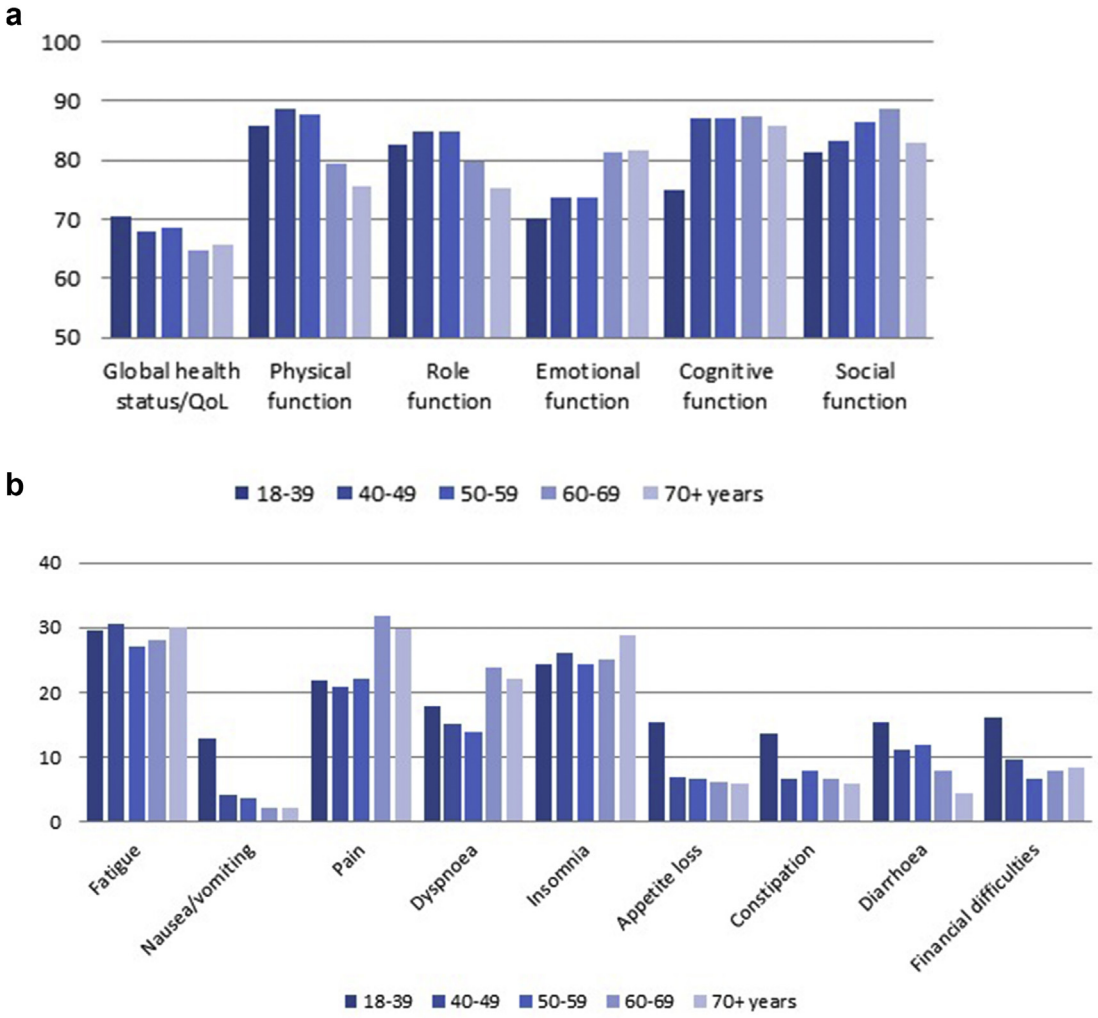

Fig. 2. a) German EORTC QLQ-C30 general population normative data for global health status/quality of life and function subscales for men by age group. b) German EORTC QLQ-C30 general population normative data for symptom subscales and items for men by age group. EORTC QLQ-C30, European Organisation for Research and Treatment of Cancer core questionnaire.

contrast, normative data reported by Waldmann *et al.* [16] are largely in line with our data, with the only difference >5 points seen in emotional functioning in women (mean score 6.1 points higher/better in our data) and dyspnoea in men (5.0 points higher/worse dyspnoea score in our data) [16]. It appears that especially male respondents in the studies by Schwarz and Hinz [15] and Hinz *et al.* [17] reported high functioning scores with >90 points in all but the emotional function scale and comparatively low symptom burden compared with our and the Waldmann *et al.* [16] data. Hence, the former samples may have consisted of respondents who may have been too healthy to be representative of the German general population. To substantiate this notion, we compared our sample with national data collected as part of the German Health Update (GEDA) study, a large-scale health monitoring study in Germany with more than 20,000 participants [29,30]. Owing to space constraints, we cannot show all details of this comparison, but in summary, we found that the GEDA 12-month prevalence data of, for example, asthma (6.2% GEDA; 7.0% in our data set) and depression (8.1%; 8.2%) in the adult population are remarkably similar, while chronic obstructive pulmonary disease prevalence is even lower in our data set (5.8%; 3.6%) [29,30]. Sixty-three percent of our sample reported at least one health condition/disease (lifetime prevalence), with presence of a health condition being clearly associated with worse functioning/higher symptom burden (data not shown). We, therefore, believe that our sampling should have captured a more representative sample of the German general population compared with previously published samples.

This study has strengths and limitations. To our knowledge, the cross-country norm data project funded by the EORTC Quality of Life Group is the largest study worldwide to generate general population norm data for the EORTC QLQ-C30. These were generated across 15 countries in Europe and North America, thereby applying a common methodology. To achieve

quotas across countries, we subcontracted data collection to one of the largest panel research companies worldwide that ensure representativeness of their online panels. However, while we believe the sampling strategy via GfK's online panels is one of the study's strengths, there are limitations. Although internet access and usage has substantially increased over the last two decades, with 88% of German households having internet access, older citizens use the internet substantially less frequently than younger generations. While generally well over 95% of the German population between 10 and 64 years uses the internet, this percentage drops to 75% for men and 60% for women, respectively, in the age group ≥65 years [31]. Therefore, it cannot be ruled out that especially the oldest age group in our sample may not be representative of the general population of that age group. However, self-selection bias is not unique to our study but is a general concern in population-based surveys. For example, there is evidence suggesting that participants in health surveys report better health-directed activities and health status overall than those who do not participate [32,33]; however, another study found that self-selection hardly influenced scores [34]. In summary, while online surveys are not free of bias, alternative data collection methods come at the expense of other biases. For example, a serious concern in the collection of population data using alternative sampling techniques is non-response bias [33], which was not as much of an issue in our online panel [19]. Furthermore, one strategy to overcome a potential limitation of coverage error by means of quota sampling was to weight our data by Germany's sex/age distribution using the population distribution statistics published by the United Nations [23]. Therefore, while bias cannot be ruled out in any population health survey, the EORTC QLQ-C30 general population norm data for Germany presented herein are the best available data to date.

5. Conclusions

This study presents updated EORTC QLQ-C30 general population norm data for Germany assessed via an online panel. The current data show some discrepancies with earlier publications on norm data from the general German population. Following our earlier discussion, we are confident that our data collection as carried out by GfK yielded high-quality data. Furthermore, the data presented herein were gathered as part of a multinational study which comes with the advantage of enabling valid inter-country comparisons. In conclusion, this study presents updated EORTC QLQ-C30 general population norm data for Germany that we recommend using for comparative purposes with data obtained from patients with cancer, in particular also for use in multinational studies.

Ethical statement

Ethical approval was not sought as this study is solely based on panel research data. As opposed to medical research where medical professional codes of conduct apply, there is widespread agreement that health research involving volunteers from the general population is not subject to ethical approval. Both the European Pharmaceutical Market Research Association (EphMRA) and the NHS Health Research Authority specify that this type of research does not require ethical approval as long as the research conforms to ethical guidelines. Our online survey was carried out by the panel research company GfK SE which is a member of EphMRA. The multinational survey conformed to the required ethical standards by obtaining informed consent from all participants and collecting data completely anonymously. Any identification of the respondents through the authors is impossible.

Conflict of interest statement

None declared.

Acknowledgements

This research was funded by the European Organisation for Research and Treatment of Cancer (EORTC) Quality of Life Group (grant number 001 2015).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejca.2020.06.002.

References

[1] Blazeby JM, Avery K, Sprangers M, Pikhart H, Fayers P, Donovan J. Health-related quality of life measurement in randomized clinical trials in surgical oncology. J Clin Oncol 2006; 24(19):3178–86.
[2] Kluetz PG, O'Connor DJ, Soltys K. Incorporating the patient experience into regulatory decision making in the USA, Europe, and Canada. Lancet Oncol 2018;19(5):e267–74.
[3] Shields AL, Hao Y, Krohe M, Yaworsky A, Mazar I, Foley C, et al. Patient-reported outcomes in oncology drug labeling in the United States: a framework for navigating early challenges. Am Health Drug Benefits 2016;9(4):188–97.
[4] Fayers P, Aaronson N, Bjordal K, Groenvold M, Curran D, Bottomley A, on behalf of the EORTC Quality of Life Group. The EORTC QLQ-C30 scoring manual. 3rd ed. Brussels: European Organisation for Research and Treatment of Cancer; 2001.
[5] Scott NW, Fayers PM, Aaronson NK, Bottomley A, de Graeff A, Groenvold M, et al., on behalf of the EORTC Quality of Life Group. In: EORTC QLQ-C30 reference values. Brussels, Belgium: E.H. Quality of Life Department.; 2008.
[6] Klee M, Groenvold M, Machin D. Quality of life of Danish women: population-based norms of the EORTC QLQ-C30. Qual Life Res 1997;6(1):27–34.

[7] Juul T, Petersen MA, Holzner B, Laurberg S, Christensen P, Gronvold M. Danish population-based reference data for the EORTC QLQ-C30: associations with gender, age and morbidity. Qual Life Res 2014;23(8):2183–93.
[8] van de Poll-Franse LV, Mols F, Gundy CM, Creutzberg CL, Nout RA, Verdonck-de Leeuw IM, et al. Normative data for the EORTC QLQ-C30 and EORTC-sexuality items in the general Dutch population. Eur J Canc 2011;47(5):667–75.
[9] Mols F, Husson O, Oudejans M, Vlooswijk C, Horevoorts N, van de Poll-Franse LV. Reference data of the EORTC QLQ-C30 questionnaire: five consecutive annual assessments of approximately 2000 representative Dutch men and women. Acta Oncol 2018:1–11.
[10] Hjermstad MJ, Fayers PM, Bjordal K, Kaasa S. Using reference data on quality of life – the importance of adjusting for age and gender, exemplified by the EORTC QLQ-C30 (+3). Eur J Canc 1998;34(9):1381–9.
[11] Hjermstad MJ, Fayers PM, Bjordal K, Kaasa S. Health-related quality of life in the general Norwegian population assessed by the European organization for research and treatment of cancer core quality-of-life questionnaire: the QLQ=C30 (+ 3). J Clin Oncol 1998;16(3):1188–96.
[12] Velenik V, Secerov-Ermenc A, But-Hadzic J, Zadnik V. Health-related quality of life assessed by the EORTC QLQ-C30 questionnaire in the general Slovenian population. Radiol Oncol 2017; 51(3):342–50.
[13] Michelson H, Bolund C, Nilsson B, Brandberg Y. Health-related quality of life measured by the EORTC QLQ-C30–reference values from a large sample of Swedish population. Acta Oncol 2000;39(4):477–84.
[14] Derogar M, van der Schaaf M, Lagergren P. Reference values for the EORTC QLQ-C30 quality of life questionnaire in a random sample of the Swedish population. Acta Oncol 2012;51(1):10–6.
[15] Schwarz R, Hinz A. Reference data for the quality of life questionnaire EORTC QLQ-C30 in the general German population. Eur J Canc 2001;37(11):1345–51.
[16] Waldmann A, Schubert D, Katalinic A. Normative data of the EORTC QLQ-C30 for the German population: a population-based survey. PloS One 2013;8(9):e74149.
[17] Hinz A, Singer S, Brahler E. European reference values for the quality of life questionnaire EORTC QLQ-C30: results of a German investigation and a summarizing analysis of six European general population normative studies. Acta Oncol 2014; 53(7):958–65.
[18] Arndt V, Koch-Gallenkamp L, Jansen L, Bertram H, Eberle A, Holleczek B, et al. Quality of life in long-term and very long-term cancer survivors versus population controls in Germany. Acta Oncol 2017:1–8.
[19] Nolte S, Liegl G, Petersen MA, Aaronson NK, Costantini A, Fayers PM, et al., on behalf of the EORTC Quality of Life Group. General population normative data for the EORTC QLQ-C30 health-related quality of life questionnaire based on 15,386 persons across 13 European countries, Canada and the Unites States. Eur J Canc 2019;107:153–63.
[20] Petersen MA, Aaronson NK, Arraras JI, Chie WC, Conroy T, Costantini A, et al., on behalf of the EORTC Quality of Life Group. The EORTC CAT Core-The computer adaptive version of the EORTC QLQ-C30 questionnaire. Eur J Canc 2018;100: 8–16.
[21] Liegl G, Petersen MA, Groenvold M, Aaronson NK, Costantini A, Fayers PM, et al., on behalf of the EORTC Quality of Life Group. Establishing the European Norm for the health-related quality of life domains of the computer-adaptive test EORTC CAT Core. Eur J Canc 2019;107:133–41.
[22] Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. J Clin Oncol 1998;16(1):139–44.
[23] United Nations Department of Economic and Social Affairs Population Division. World population Prospects: the 2017 revision. DVD Edition; 2017.
[24] Federal statistical office and the statistical offices of the länder. *Result of the 2011 Census at the reference date 9th may 2011*. 2014.
[25] Liu H, Cella D, Gershon R, Shen J, Morales LS, Riley W, et al. Representativeness of the patient-reported outcomes measurement information system internet panel. J Clin Epidemiol 2010; 63(11):1169–78.
[26] Crumley JJ, Stetler CA, Horhota M. Examining the relationship between subjective and objective memory performance in older adults: a meta-analysis. Psychol Aging 2014;29(2):250–63.
[27] Wills TA. Downward comparison principles in social psychology. Psychol Bull 1981;90(2):245–71.
[28] Lange C, Manz K, Kuntz B. Alcohol consumption among adults in Germany: heavy episodic drinking. Journal of Health Monitoring 2017;2(2):71–7.
[29] Steppuhn H, Buda S, Wienecke A, Kraywinkel K, Tolksdorf K, Haberland J, et al. Time trends in incidence and mortality of respiratory diseases of high public health relevance in Germany. Epidemiologie und Gesundheitsberichterstattung. Robert Koch-Institut; 2017.
[30] Thom J, Kuhnert R, Born S, Hapke U. 12-month prevalence of self-reported medical diagnoses of depression in Germany. Epidemiologie und Gesundheitsberichterstattung. Robert Koch-Institut; 2017.
[31] Destatis. Survey on the private use of information and communication technologies (ICT). In: Fachserie 15, Reihe 4. Federal Statistical Office of Germany; 2019.
[32] Van Loon AJ, Tijhuis M, Picavet HS, Surtees PG, Ormel J. Survey non-response in The Netherlands: effects on prevalence estimates and associations. Ann Epidemiol 2003;13(2):105–10.
[33] Cheung KL, Ten Klooster PM, Smit C, de Vries H, Pieterse ME. The impact of non-response bias due to sampling in public health studies: a comparison of voluntary versus mandatory recruitment in a Dutch national survey on adolescent health. BMC Publ Health 2017;17(1):276.
[34] Søgaard AJ, Selmer R, Bjertness E, Thelle D. The Oslo Health Study: the impact of self-selection in a large, population-based survey. Int J Equity Health 2004;3(1). 3-3.

3 Discussion

Research aimed at improving the quality of PRO assessment as well as the actual application of PROs in both clinical practice and clinical research have grown exponentially over the last few decades. Many PRO-specific guidelines and standards now exist that – if followed closely – should greatly enhance the quality of the PRO data collected. In essence, if high quality measures with robust validity evidence are applied, users and other stakeholders of PRO data can be reasonably confident that the inferences that are drawn from these data are valid (Hawkins et al., 2021; Hawkins et al., 2018; Weinfurt, 2021, 2022).

Two international PRO measurement systems were introduced here that are exemplars of rigorous PRO measure development and validation processes. First, over the last decade the U.S.-based PROMIS initiative has developed and evaluated a range of PRO measures for the assessment of mental, physical and social constructs that are aimed to be applied across diseases. PROMIS is the first initiative of its kind that introduced a measurement system for PROs that is based on modern test theory methods that allow for the application of CAT (Cella et al., 2010; Fries et al., 2005). Second, the PRO measurement system of the EORTC is a European initiative that was originally based on static PRO instruments for the assessment of cancer patients' QoL (Aaronson et al., 1993). More recently, the EORTC has also released a CAT version of the EORTC QLQ-C30 (Petersen et al., 2018; Petersen et al., 2020). Similar to PROMIS, the EORTC has established standardised procedures that are applied during the development of their cancer-specific modules (Wheelwright et al., 2021) as well as during the numerous translation and cultural adaptation projects (Kuliś et al., 2017). In summary, both initiatives have contributed greatly to the standardisation of PRO assessment and they have raised the profile of PRO data as an important source of information for medical and health policy decision making (Porter et al., 2016).

The included articles made critical contributions to both of these prominent efforts of the international PRO community. The first study contributed to the PROMIS initiative by translating and culturally adapting the PROMIS Physical Function item bank into German and by subsequently evaluating its psychometric properties. Since the initial paper, a more thorough investigation of this item bank has been undertaken, including DIF analyses to explore the psychometric performance of the item bank in different disease groups (Liegl et al., 2020). Owing to this DFG-funded work, the project has made the PROMIS Physical

Function item bank available to the German-speaking community after confirming instrument equivalence across language versions and it also contributed more generally to the research question about dimensionality of the item bank (Liegl et al., 2018; Liegl et al., 2020). The second PROMIS project is an exemplar of a standardised evaluation of the psychometric properties of a PRO measure using CTT methods. Being part of a larger collaborative effort that was aimed at comparing three test-theoretical approaches for the evaluation of the PROMIS Depression item bank (Cleanthous et al., 2019; Nolte, Coon, et al., 2019; Stover et al., 2019), the project also needs to be seen in this broader context as it made a crucial methodological contribution to the wider PRO community by exploring the differences and commonalities between CTT, IRT and RMT (Bjorner, 2019). In contrast to above PROMIS projects, the EORTC-funded work on norm data contributed to the interpretation of both the EORTC CAT Core and the EORTC QLQ-C30. By establishing general population norm data for both instruments, the interpretability of scores obtained from these two measures has been enhanced greatly and – given that a consistent sampling method was applied across countries – the updated norm data are also suitable for inter-country comparisons (Liegl et al., 2019; Nolte, Liegl, et al., 2019; Nolte et al., 2020). Since the three EORTC articles were published, additional national general population norm data have been released (Arraras et al., 2021; Lehmann et al., 2020; Pilz et al., 2022), with further national norm data publications being underway. In addition, other research groups have also started making use of the norm data (Karsten et al., 2022; Ludwig et al., 2020; Schwartz et al., 2021).

Above projects made crucial contributions to the international PRO community, and both the PROMIS and EORTC measurement systems are popular and used widely across the world; however, they compete with a plethora of PRO measures that have been developed and released over the last few decades. To get a sense of the number of PRO instruments available, the Mapi Research Trust offers a great resource for users of PROs. The Trust's database PROQOLID is a library of COA measures – many of which are PRO measures – that currently contains information on over 5,100 COA measures worldwide. As can be seen on the Trust's website (https://eprovide.mapi-trust.org/), the sheer volume of measures that can be chosen from make the field of PRO/COA assessment rather confusing, especially in those contexts where various PRO instruments are available to assess the same latent construct. For example, Wahl et al. (2011) identified in excess of 100 PRO measures for the

assessment of depression alone, making PRO measure selection a real challenge, especially for less experienced PRO users.

In view of the plethora of PRO measures available that challenge comparison of PRO scores obtained from different instruments, modern test theory methods offer the opportunity to link these scores (Schalet et al., 2015). Such linking can be achieved by calibrating items that measure the same latent construct on the same scale, taking the standardisation of PRO measurement to the next level. This item calibration process enables the comparison of scores derived from any subset of items from within the same but also across different measures. To date, this approach has been applied by various research groups that have linked different PRO measures via a common metric, for example, for the assessment of anxiety, depression, pain or physical function (Choi et al., 2014; Cook et al., 2015; Kaat et al., 2017; Kaat et al., 2018; Liegl et al., 2016; Oude Voshaar et al., 2019; Schalet et al., 2014; Schalet et al., 2015; Wahl et al., 2014). In addition, the website www.common-metrics.org provided by Fischer and Rose (2016) offers a convenient way of linking scores from different PRO measures (Fischer & Rose, 2016). Albeit unrealistic, the idea of creating a common metric for various constructs could yet be taken further again by moving from instrument-based to construct-based PRO assessment (Kaat et al., 2018; Liegl et al., 2020). As described above, it is not only possible to calibrate items from different PRO measures on the same metric and make PRO scores comparable but it is also possible that this new item pool could become a new PRO measure in its own right. This measure could be continuously improved by only retaining those items with good item parameters and subsequently adding new items, for example, in order to extend the measurement range which was done for the PROMIS Physical Function item bank (Kaat et al., 2019). This would then constitute a real move to construct-based PRO assessment.

Despite the appeal of construct-based PRO assessment that is independent of specific PRO instruments, it is unrealistic that this will ever be implemented as there are too many PRO measures that are already used widely. Also, PRO measure developers are unlikely to be willing to give up their own PRO measure. Therefore, standardisation of PRO assessment reaches its limits despite methods being available that could turn instrument-independent PRO assessment into reality. Consequently, it remains vital that users of PROs are not only aware of but also follow the many PRO-specific guidance documents, standards as well as recommendations aimed at improving the data quality obtained via PROs. As presented in

the Introduction, central publications in this area relate to measure development and content validation (Patrick et al., 2011a, 2011b), design and selection of PRO measures (Reeve et al., 2013), translation/cultural adaptation of PRO measures (Wild et al., 2005) and definitions of key attributes of PRO measures (Aaronson et al., 2002; Mokkink et al., 2010), whilst others set out to standardise the statistical analysis of PRO data (Coens et al., 2020).

Above advancements have greatly contributed to the standardisation of PRO measurement over the last few years and decades. If PRO users closely follow the many published PRO-specific guidelines, it should greatly enhance PRO data quality. However, a last vital step is data interpretation, i.e., assuming robust validity evidence has been gathered, inferences that are drawn from these data are only valid if PRO users know how to interpret the data. Even though data interpretation is not directly the focus of PRO standardisation efforts, various approaches to interpret PRO data have been published that implicitly contribute to PRO standardisation. As implemented by the EORTC, one possible approach to interpret quantitative data is the comparison of patients' scores with reference values obtained from patients with comparable conditions (Scott et al., 2008) or with norm data collected from a representative sample of a country's general population (Nolte, Liegl, et al., 2019). As it is rather laborious and costly to obtain reference values and/or general population norm data, these are typically only available for PRO measures that were developed as part of larger initiatives, such as the EORTC and the PROMIS suites of PRO instruments (Jensen et al., 2017; Liegl et al., 2019; Nolte, Liegl, et al., 2019; Rose et al., 2014; Scott et al., 2008). Alternatively, statistical significance testing may be applied to explore differences at the group level. However, in light of the limitations of statistical significance testing that is highly dependent on the sample size and has little intrinsic meaning regarding clinical significance of the findings, Jaeschke et al. (1989) introduced the concept of minimal clinically important differences (MCID) as an alternative approach to interpret change over time.

The seminal papers by Jaeschke et al. (1989) and Lydick and Epstein (1993) who introduced anchor- and distribution-based methods to interpret QoL data laid the groundwork for a large amount of research undertaken in the area of MCID. The idea behind MCID is to improve the interpretability of PRO scores by defining thresholds a priori that define a level of change at the individual or at the group level at which it is deemed clinically meaningful (Jaeschke et al., 1989). When deriving these thresholds, it is crucial to differentiate between individual within-person change over time, group-level change over time and between-

group differences, and there are various approaches to derive these thresholds (Coon & Cappelleri, 2016; Coon & Cook, 2018). Especially since the 2009 FDA PRO Guidance for Industry where the use of a responder threshold was promoted, i.e., a threshold for interpretation of within-person change (FDA, 2009), publications on thresholds to define clinical meaningfulness have grown exponentially. To get an overview of these cut-off points and help PRO users navigate through the field of minimal important differences (MID)[2], researchers at McMaster University initiated PROMID, i.e., a MID database for PRO measures (https://promid.mcmaster.ca/) (Johnston et al., 2015). In their accompanying publication, they describe that the PROMID inventory contains 5,324 unique MID estimates for 526 distinct PRO measures (Carrasco-Labra et al., 2021). While PRO users need to be cautious regarding the quality of the MID estimates, for which the McMaster group has developed a credibility instrument to judge the quality of MIDs derived from anchor-based methods (Devji et al., 2020), PROMID is a new and important effort towards standardising PRO data interpretation that supports PRO users in understanding the data they obtain from their patients.

The many PRO-specific guidance publications introduced above taken together with the various approaches to improving the interpretation of PRO data greatly contribute to raising the level of PRO data quality and subsequently the accumulation of high quality validity evidence that is necessary to draw valid inferences from these data (Hawkins et al., 2021; Hawkins et al., 2018; Weinfurt, 2021, 2022). Despite all of these advancements, however, it remains that there is a large volume of PRO measures available, many of which have not been developed applying state-of-the-art methods, have not undergone thorough psychometric evaluation, and neither reference/general population norm data nor MID thresholds are available to aid data interpretation. In the process of increased expectations at PRO data quality not only from regulatory bodies, such as the FDA, but also from the broader PRO research community, it is expected that there will be a consolidation phase regarding PRO measures. Therefore, those measures are most likely to survive that fulfil the various PRO quality criteria and that offer a range of options for data interpretation, including reference values, general population norm data and thresholds that signify

[2] Since the introduction of the term MCID by Jaeschke et al. (1989), a large volume of articles on this topic have been published, including different abbreviations and definitions, different methods and different application (Coon & Cappelleri, 2016). The use of MID here is chosen for the sole purpose of describing a threshold for clinical significance.

clinically meaningful change or difference. Both the PROMIS and the EORTC suites of PRO instruments are likely PRO measurement systems that will contribute to the anticipated consolidation phase of PRO measures with the ultimate goal of achieving standardised PRO measurement.

4 Summary and Outlook

Initiatives contributing to the standardisation of PRO assessment are well underway and various publications provide clear recommendations regarding key aspects of standardised PRO assessment, including recommendations in the context of development/selection and content validation of PRO measures, translation and cultural adaptation, key attributes of PRO measures and statistical analysis of PRO data (Aaronson et al., 2002; Mokkink et al., 2010; Patrick et al., 2011a, 2011b; Reeve et al., 2013; Wild et al., 2005). In addition, various efforts have been undertaken with the aim to facilitate PRO data interpretation, including publications on reference values (Scott et al., 2008), general population norm data (Nolte, Liegl, et al., 2019) as well as MID thresholds for a large range of PRO measures (Carrasco-Labra et al., 2021). Finally, several guidelines have been published that provide standards on how to include PROs in clinical trial protocols, publications and reports (Basch & Leahy, 2019; Calvert et al., 2013; Calvert et al., 2018; Coens et al., 2020). Hence, provided above recommendations are followed, PRO users should be able to produce high quality PRO measures that generate robust validity evidence that in turn should lead to high quality data. High quality PRO data are vital for the various stakeholders of PROs to confidently use these data as a base for medical and health policy decision making (Hawkins et al., 2021; Hawkins et al., 2018; Weinfurt, 2021, 2022). Despite these advancements, it remains that there is currently a large volume of PRO measures available with varying quality. Therefore, it is expected that there will soon be a consolidation phase, with those PRO measures likely to survive that fulfil PRO quality criteria as introduced above and that offer various options for data interpretation. The work presented here made critical contributions to standardised PRO measurement in the context of translation and cultural adaption of a PRO measure, the psychometric evaluation of PRO measures as well as PRO data interpretation. All projects were carried out in the context of two of the main PRO initiatives worldwide. It is expected that both the PROMIS and EORTC suites of PRO measures will continue to grow, consolidate the PRO field and continue to be a benchmark for standardised PRO assessment, validation and interpretation.

5 References

Aaronson, N., Alonso, J., Burnam, A., Lohr, K. N., Patrick, D. L., Perrin, E., & Stein, R. E. (2002). Assessing health status and quality-of-life instruments: attributes and review criteria. *Qual Life Res*, *11*(3), 193-205. https://doi.org/10.1023/a:1015291021312

Aaronson, N. K., Acquadro, C., Alonso, J., Apolone, G., Bucquet, D., Bullinger, M., Bungay, K., Fukuhara, S., Gandek, B., Keller, S., & et al. (1992). International Quality of Life Assessment (IQOLA) Project. *Qual Life Res*, *1*(5), 349-351. https://doi.org/10.1007/bf00434949

Aaronson, N. K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N. J., Filiberti, A., Flechtner, H., Fleishman, S. B., de Haes, J. C., & et al. (1993). The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst*, *85*(5), 365-376.

Ahmed, S., Berzon, R. A., Revicki, D. A., Lenderking, W. R., Moinpour, C. M., Basch, E., Reeve, B. B., & Wu, A. W. (2012). The use of patient-reported outcomes (PRO) within comparative effectiveness research: implications for clinical practice and health care policy. *Med Care*, *50*(12), 1060-1070. https://doi.org/10.1097/MLR.0b013e318268aaff

Andrich, D. (2011). Rating scales and Rasch measurement. *Expert Rev Pharmacoecon Outcomes Res*, *11*(5), 571-585. https://doi.org/10.1586/erp.11.59

Arraras, J. I., Nolte, S., Liegl, G., Rose, M., Manterola, A., Illarramendi, J. J., Zarandona, U., Rico, M., Teiejria, L., Asin, G., Hernandez, I., Barrado, M., Vera, R., Efficace, F., & Giesinger, J. M. (2021). General Spanish population normative data analysis for the EORTC QLQ-C30 by sex, age, and health condition. *Health Qual Life Outcomes*, *19*(1), 208. https://doi.org/10.1186/s12955-021-01820-x

Atherton, P. J., & Sloan, J. A. (2006). Rising importance of patient-reported outcomes. *Lancet Oncol*, *7*(11), 883-884. https://doi.org/10.1016/s1470-2045(06)70914-7

Basch, E. (2017). Patient-Reported Outcomes - Harnessing Patients' Voices to Improve Clinical Care. *N Engl J Med*, *376*(2), 105-108. https://doi.org/10.1056/NEJMp1611252

Basch, E., & Leahy, A. B. (2019). Reporting Standards for Patient-Reported Outcomes in Clinical Trial Protocols and Publications. *J Natl Cancer Inst*, *111*(11), 1116-1117. https://doi.org/10.1093/jnci/djz047

Bjorner, J. B. (2019). State of the psychometric methods: comments on the ISOQOL SIG psychometric papers. *J Patient Rep Outcomes*, *3*(1), 49. https://doi.org/10.1186/s41687-019-0134-1

Calvert, M., Blazeby, J., Altman, D. G., Revicki, D. A., Moher, D., & Brundage, M. D. (2013). Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. *JAMA*, *309*(8), 814-822. https://doi.org/10.1001/jama.2013.879

Calvert, M., Kyte, D., Mercieca-Bebber, R., Slade, A., Chan, A. W., King, M. T., Hunn, A., Bottomley, A., Regnault, A., Chan, A. W., Ells, C., O'Connor, D., Revicki, D., Patrick, D., Altman, D., Basch, E., Velikova, G., Price, G., Draper, H., . . . Groves, T. (2018). Guidelines for Inclusion of Patient-Reported Outcomes in Clinical Trial Protocols: The SPIRIT-PRO Extension. *JAMA*, *319*(5), 483-494. https://doi.org/10.1001/jama.2017.21903

Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clin Ther*, *36*(5), 648-662. https://doi.org/10.1016/j.clinthera.2014.04.006

Carrasco-Labra, A., Devji, T., Qasim, A., Phillips, M. R., Wang, Y., Johnston, B. C., Devasenapathy, N., Zeraatkar, D., Bhatt, M., Jin, X., Brignardello-Petersen, R., Urquhart, O., Foroutan, F., Schandelmaier, S., Pardo-Hernandez, H., Hao, Q., Wong, V., Ye, Z., Yao, L., . . . Guyatt, G. H. (2021). Minimal important difference estimates for patient-reported outcomes: A systematic survey. *J Clin Epidemiol*, *133*, 61-71. https://doi.org/10.1016/j.jclinepi.2020.11.024

Cella, D., Choi, S. W., Condon, D. M., Schalet, B., Hays, R. D., Rothrock, N. E., Yount, S., Cook, K. F., Gershon, R. C., Amtmann, D., DeWalt, D. A., Pilkonis, P. A., Stone, A. A., Weinfurt, K., & Reeve, B. B. (2019). PROMIS(®) Adult Health Profiles: Efficient Short-Form Measures of Seven Health Domains. *Value Health*, *22*(5), 537-544. https://doi.org/10.1016/j.jval.2019.02.004

Cella, D., Gershon, R., Lai, J. S., & Choi, S. (2007). The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. *Qual Life Res*, *16 Suppl 1*, 133-141. https://doi.org/10.1007/s11136-007-9204-6

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D., Choi, S., Cook, K., Devellis, R., DeWalt, D., Fries, J. F., Gershon, R., Hahn, E. A., Lai, J. S., Pilkonis, P., Revicki, D., . . . Hays, R. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. *J Clin Epidemiol*, *63*(11), 1179-1194. https://doi.org/10.1016/j.jclinepi.2010.04.011

Choi, S. W., Schalet, B., Cook, K. F., & Cella, D. (2014). Establishing a common metric for depressive symptoms: linking the BDI-II, CES-D, and PHQ-9 to PROMIS depression. *Psychol Assess*, *26*(2), 513-527. https://doi.org/10.1037/a0035768

Cleanthous, S., Barbic, S. P., Smith, S., & Regnault, A. (2019). Psychometric performance of the PROMIS® depression item bank: a comparison of the 28- and 51-item versions using Rasch measurement theory. *J Patient Rep Outcomes*, *3*(1), 47. https://doi.org/10.1186/s41687-019-0131-4

Coens, C., Pe, M., Dueck, A. C., Sloan, J., Basch, E., Calvert, M., Campbell, A., Cleeland, C., Cocks, K., Collette, L., Devlin, N., Dorme, L., Flechtner, H. H., Gotay, C., Griebsch, I., Groenvold, M., King, M., Kluetz, P. G., Koller, M., . . . Bottomley, A. (2020). International standards for the analysis of quality-of-life and patient-reported outcome endpoints in cancer randomised controlled trials: recommendations of the SISAQOL Consortium. *Lancet Oncol*, *21*(2), e83-e96. https://doi.org/10.1016/s1470-2045(19)30790-9

Cook, K. F., Schalet, B. D., Kallen, M. A., Rutsohn, J. P., & Cella, D. (2015). Establishing a common metric for self-reported pain: linking BPI Pain Interference and SF-36 Bodily Pain Subscale scores to the PROMIS Pain Interference metric. *Qual Life Res*, *24*(10), 2305-2318. https://doi.org/10.1007/s11136-015-0987-6

Coon, C. D., & Cappelleri, J. C. (2016). Interpreting Change in Scores on Patient-Reported Outcome Instruments. *Ther Innov Regul Sci*, *50*(1), 22-29. https://doi.org/10.1177/2168479015622667

Coon, C. D., & Cook, K. F. (2018). Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. *Qual Life Res*, *27*(1), 33-40. https://doi.org/10.1007/s11136-017-1616-3

De Brabander, B., & Gerits, P. (1999). Chronic and acute stress as predictors of relapse in primary breast cancer patients. *Patient Educ Couns*, *37*(3), 265-272.

Degner, L. F., & Sloan, J. A. (1995). Symptom distress in newly diagnosed ambulatory cancer patients and as a predictor of survival in lung cancer. *J Pain Symptom Manage*, *10*(6), 423-431. https://doi.org/10.1016/0885-3924(95)00056-5

Derogar, M., van der Schaaf, M., & Lagergren, P. (2012). Reference values for the EORTC QLQ-C30 quality of life questionnaire in a random sample of the Swedish population. *Acta Oncol*, *51*(1), 10-16. https://doi.org/10.3109/0284186x.2011.614636

DeVellis, R. F. (2006). Classical test theory. *Med Care*, *44*(11 Suppl 3), S50-59. https://doi.org/10.1097/01.mlr.0000245426.10853.30

Devji, T., Carrasco-Labra, A., Qasim, A., Phillips, M., Johnston, B. C., Devasenapathy, N., Zeraatkar, D., Bhatt, M., Jin, X., Brignardello-Petersen, R., Urquhart, O., Foroutan, F., Schandelmaier, S., Pardo-Hernandez, H., Vernooij, R. W., Huang, H., Rizwan, Y., Siemieniuk, R., Lytvyn, L., . . . Guyatt, G. H. (2020). Evaluating the credibility of anchor based estimates of minimal important differences for patient reported outcomes: instrument development and reliability study. *BMJ*, *369*, m1714. https://doi.org/10.1136/bmj.m1714

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates.

European Medicines Agency. (2005). *Reflection paper on the regulatory guidance for the use of health-related quality of life (HRQL) measures in the evaluation of medicinal products*. European Medicines Agency website. Retrieved 23 Apr 2022 from https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-regulatory-guidance-use-healthrelated-quality-life-hrql-measures-evaluation_en.pdf

FDA. (2006). Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. *Health Qual Life Outcomes*, *4*, 79. https://doi.org/10.1186/1477-7525-4-79

FDA. (2009). *Guidance for industry. Patient-reported outcome measures: use in medical product development to support labeling claims*. U.S. Food and Drug Administration website. Retrieved 23 Apr 2022 from http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf

FDA. (2020). *Patient-Focused Drug Development: Collecting Comprehensive and Representative Input*. U.S. Food and Drug Administration website. Retrieved 23 Apr 2022 from https://www.fda.gov/media/139088/download

FDA. (2022). *Patient-Focused Drug Development: Methods to Identify What Is Important to Patients*. U.S. Food and Drug Administration website. Retrieved 23 Apr 2022 from https://www.fda.gov/media/131230/download

Fischer, H. F., & Rose, M. (2016). www.common-metrics.org: a web application to estimate scores from different patient-reported outcome measures on a common scale [journal

article]. *BMC Medical Research Methodology*, *16*(1), 142. https://doi.org/10.1186/s12874-016-0241-0

Frank, L., Basch, E., & Selby, J. V. (2014). The PCORI perspective on patient-centered outcomes research. *JAMA*, *312*(15), 1513-1514. https://doi.org/10.1001/jama.2014.11100

Fries, J. F., Bruce, B., & Cella, D. (2005). The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes. *Clin Exp Rheumatol*, *23*(5 Suppl 39), S53-57.

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., Bhaumik, D. K., Stover, A., Bock, R. D., & Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatr Serv*, *59*(4), 361-368. https://doi.org/10.1176/appi.ps.59.4.361

Gotay, C. C., Kawamoto, C. T., Bottomley, A., & Efficace, F. (2008). The prognostic significance of patient-reported outcomes in cancer clinical trials. *J Clin Oncol*, *26*(8), 1355-1363. https://doi.org/10.1200/jco.2007.13.3439

Groenvold, M., Petersen, M. A., Idler, E., Bjorner, J. B., Fayers, P. M., & Mouridsen, H. T. (2007). Psychological distress and fatigue predicted recurrence and survival in primary breast cancer patients. *Breast Cancer Res Treat*, *105*(2), 209-219. https://doi.org/10.1007/s10549-006-9447-x

Hawkins, M., Elsworth, G. R., Nolte, S., & Osborne, R. H. (2021). Validity arguments for patient-reported outcomes: justifying the intended interpretation and use of data. *J Patient Rep Outcomes*, *5*(1), 64. https://doi.org/10.1186/s41687-021-00332-y

Hawkins, M., Elsworth, G. R., & Osborne, R. H. (2018). Application of validity theory and methodology to patient-reported outcome measures (PROMs): building an argument for validity. *Qual Life Res*, *27*(7), 1695-1710. https://doi.org/10.1007/s11136-018-1815-6

Hinz, A., Singer, S., & Brahler, E. (2014). European reference values for the quality of life questionnaire EORTC QLQ-C30: Results of a German investigation and a summarizing analysis of six European general population normative studies. *Acta Oncol*, *53*(7), 958-965. https://doi.org/10.3109/0284186x.2013.879998

Jaeschke, R., Singer, J., & Guyatt, G. (1989). Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials*, *10*, 407-415.

Jensen, R. E., Potosky, A. L., Moinpour, C. M., Lobo, T., Cella, D., Hahn, E. A., Thissen, D., Smith, A. W., Ahn, J., Luta, G., & Reeve, B. B. (2017). United States Population-Based Estimates of Patient-Reported Outcomes Measurement Information System Symptom and Functional Status Reference Values for Individuals With Cancer. *J Clin Oncol*, *35*(17), 1913-1920. https://doi.org/10.1200/jco.2016.71.4410

Johnston, B. C., Ebrahim, S., Carrasco-Labra, A., Furukawa, T. A., Patrick, D. L., Crawford, M. W., Hemmelgarn, B. R., Schunemann, H. J., Guyatt, G. H., & Nesrallah, G. (2015). Minimally important difference estimates and methods: a protocol. *BMJ Open*, *5*(10), e007953. https://doi.org/10.1136/bmjopen-2015-007953

Juul, T., Petersen, M. A., Holzner, B., Laurberg, S., Christensen, P., & Gronvold, M. (2014). Danish population-based reference data for the EORTC QLQ-C30: associations with

gender, age and morbidity. *Qual Life Res*, *23*(8), 2183-2193. https://doi.org/10.1007/s11136-014-0675-y

Kaat, A. J., Buckenmaier, C. T., 3rd, Cook, K. F., Rothrock, N. E., Schalet, B. D., Gershon, R. C., & Vrahas, M. S. (2019). The expansion and validation of a new upper extremity item bank for the Patient-Reported Outcomes Measurement Information System® (PROMIS). *J Patient Rep Outcomes*, *3*(1), 69. https://doi.org/10.1186/s41687-019-0158-6

Kaat, A. J., Newcomb, M. E., Ryan, D. T., & Mustanski, B. (2017). Expanding a common metric for depression reporting: linking two scales to PROMIS(®) depression. *Qual Life Res*, *26*(5), 1119-1128. https://doi.org/10.1007/s11136-016-1450-z

Kaat, A. J., Schalet, B. D., Rutsohn, J., Jensen, R. E., & Cella, D. (2018). Physical function metric over measure: An illustration with the Patient-Reported Outcomes Measurement Information System (PROMIS) and the Functional Assessment of Cancer Therapy (FACT). *Cancer*, *124*(1), 153-160. https://doi.org/10.1002/cncr.30981

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527-535.

Karsten, M. M., Roehle, R., Albers, S., Pross, T., Hage, A. M., Weiler, K., Fischer, F., Rose, M., Kühn, F., & Blohmer, J. U. (2022). Real-world reference scores for EORTC QLQ-C30 and EORTC QLQ-BR23 in early breast cancer patients. *Eur J Cancer*, *163*, 128-139. https://doi.org/10.1016/j.ejca.2021.12.020

Kocalevent, R. D., Rose, M., Becker, J., Walter, O. B., Fliege, H., Bjorner, J. B., Kleiber, D., & Klapp, B. F. (2009). An evaluation of patient-reported outcomes found computerized adaptive testing was efficient in assessing stress perception. *J Clin Epidemiol*, *62*(3), 278-287, 287.e271-273. https://doi.org/10.1016/j.jclinepi.2008.03.003

Kuliś, D., Whittaker, C., Greimel, E., Bottomley, A., & Koller, M. (2017). Reviewing back translation reports of questionnaires: the EORTC conceptual framework and experience. *Expert Rev Pharmacoecon Outcomes Res*, *17*(6), 523-530. https://doi.org/10.1080/14737167.2017.1384316

Lehmann, J., Giesinger, J. M., Nolte, S., Sztankay, M., Wintner, L. M., Liegl, G., Rose, M., & Holzner, B. (2020). Normative data for the EORTC QLQ-C30 from the Austrian general population. *Health Qual Life Outcomes*, *18*(1), 275. https://doi.org/10.1186/s12955-020-01524-8

Liegl, G., Petersen, M. A., Groenvold, M., Aaronson, N. K., Costantini, A., Fayers, P. M., Holzner, B., Johnson, C. D., Kemmler, G., Tomaszewski, K. A., Waldmann, A., Young, T. E., Rose, M., & Nolte, S. (2019). Establishing the European Norm for the health-related quality of life domains of the computer-adaptive test EORTC CAT Core. *Eur J Cancer*, *107*, 133-141. https://doi.org/10.1016/j.ejca.2018.11.023

Liegl, G., Rose, M., Correia, H., Fischer, H. F., Kanlidere, S., Mierke, A., Obbarius, A., & Nolte, S. (2018). An initial psychometric evaluation of the German PROMIS v1.2 Physical Function item bank in patients with a wide range of health conditions. *Clin Rehabil*, *32*(1), 84-93. https://doi.org/10.1177/0269215517714297

Liegl, G., Rose, M., Knebel, F., Stengel, A., Buttgereit, F., Obbarius, A., Fischer, H. F., & Nolte, S. (2020). Using subdomain-specific item sets affected PROMIS physical function scores

differently in cardiology and rheumatology patients. *J Clin Epidemiol*, *127*, 151-160. https://doi.org/10.1016/j.jclinepi.2020.08.003

Liegl, G., Wahl, I., Berghofer, A., Nolte, S., Pieh, C., Rose, M., & Fischer, F. (2016). Using Patient Health Questionnaire-9 item parameters of a common metric resulted in similar depression scores compared to independent item response theory model reestimation. *J Clin Epidemiol*, *71*, 25-34. https://doi.org/10.1016/j.jclinepi.2015.10.006

Ludwig, H., Pönisch, W., Knop, S., Egle, A., Hinke, A., Schreder, M., Lechner, D., Hajek, R., Gunsilius, E., Petzer, A., Weisel, K., Niederwieser, D., Einsele, H., Willenbacher, W., Rumpold, H., Pour, L., Jelinek, T., Krenosz, K. J., Meckl, A., . . . Zojer, N. (2020). Quality of life in patients with relapsed/refractory multiple myeloma during ixazomib-thalidomide-dexamethasone induction and ixazomib maintenance therapy and comparison to the general population. *Leuk Lymphoma*, *61*(2), 377-386. https://doi.org/10.1080/10428194.2019.1666381

Lydick, E., & Epstein, R. (1993). Interpretation of quality of life changes. *Qual Life Res*, *2*, 221-226.

McDonald, R. (1999). *Test theory: a unified treatment*. Lawrence Erlbaum Associates.

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol*, *63*(7), 737-745. https://doi.org/10.1016/j.jclinepi.2010.02.006

Mols, F., Husson, O., Oudejans, M., Vlooswijk, C., Horevoorts, N., & van de Poll-Franse, L. V. (2018). Reference data of the EORTC QLQ-C30 questionnaire: five consecutive annual assessments of approximately 2000 representative Dutch men and women. *Acta Oncol*, 1-11. https://doi.org/10.1080/0284186x.2018.1481293

Nolte, S., Coon, C., Hudgens, S., & Verdam, M. G. E. (2019). Psychometric evaluation of the PROMIS® Depression Item Bank: an illustration of classical test theory methods. *J Patient Rep Outcomes*, *3*(1), 46. https://doi.org/10.1186/s41687-019-0127-0

Nolte, S., Liegl, G., Petersen, M. A., Aaronson, N. K., Costantini, A., Fayers, P. M., Groenvold, M., Holzner, B., Johnson, C. D., Kemmler, G., Tomaszewski, K. A., Waldmann, A., Young, T. E., & Rose, M. (2019). General population normative data for the EORTC QLQ-C30 health-related quality of life questionnaire based on 15,386 persons across 13 European countries, Canada and the Unites States. *Eur J Cancer*, *107*, 153-163. https://doi.org/10.1016/j.ejca.2018.11.024

Nolte, S., Waldmann, A., Liegl, G., Petersen, M. A., Groenvold, M., & Rose, M. (2020). Updated EORTC QLQ-C30 general population norm data for Germany. *Eur J Cancer*, *137*, 161-170. https://doi.org/10.1016/j.ejca.2020.06.002

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Oude Voshaar, M. A. H., Vonkeman, H. E., Courvoisier, D., Finckh, A., Gossec, L., Leung, Y. Y., Michaud, K., Pinheiro, G., Soriano, E., Wulfraat, N., Zink, A., & van de Laar, M. (2019). Towards standardized patient reported physical function outcome reporting: linking ten commonly used questionnaires to a common metric. *Qual Life Res*, *28*(1), 187-197. https://doi.org/10.1007/s11136-018-2007-0

Patrick, D. L., Burke, L. B., Gwaltney, C. J., Leidy, N. K., Martin, M. L., Molsen, E., & Ring, L. (2011a). Content validity--establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: part 1--eliciting concepts for a new PRO instrument. *Value Health*, *14*(8), 967-977. https://doi.org/10.1016/j.jval.2011.06.014

Patrick, D. L., Burke, L. B., Gwaltney, C. J., Leidy, N. K., Martin, M. L., Molsen, E., & Ring, L. (2011b). Content validity--establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: part 2--assessing respondent understanding. *Value Health*, *14*(8), 978-988. https://doi.org/10.1016/j.jval.2011.06.013

Petersen, M. A., Aaronson, N. K., Arraras, J. I., Chie, W. C., Conroy, T., Costantini, A., Dirven, L., Fayers, P., Gamper, E. M., Giesinger, J. M., Habets, E. J. J., Hammerlid, E., Helbostad, J., Hjermstad, M. J., Holzner, B., Johnson, C., Kemmler, G., King, M. T., Kaasa, S., . . . Groenvold, M. (2018). The EORTC CAT Core-The computer adaptive version of the EORTC QLQ-C30 questionnaire. *Eur J Cancer*, *100*, 8-16. https://doi.org/10.1016/j.ejca.2018.04.016

Petersen, M. A., Aaronson, N. K., Conroy, T., Costantini, A., Giesinger, J. M., Hammerlid, E., Holzner, B., Johnson, C. D., Kieffer, J. M., van Leeuwen, M., Nolte, S., Ramage, J. K., Tomaszewski, K. A., Waldmann, A., Young, T., Zotti, P., & Groenvold, M. (2020). International validation of the EORTC CAT Core: a new adaptive instrument for measuring core quality of life domains in cancer. *Qual Life Res*, *29*(5), 1405-1417. https://doi.org/10.1007/s11136-020-02421-9

Pilz, M. J., Gamper, E. M., Efficace, F., Arraras, J. I., Nolte, S., Liegl, G., Rose, M., & Giesinger, J. M. (2022). EORTC QLQ-C30 general population normative data for Italy by sex, age and health condition: an analysis of 1,036 individuals. *BMC Public Health*, *22*(1), 1040. https://doi.org/10.1186/s12889-022-13211-y

Porter, M. E., Larsson, S., & Lee, T. H. (2016). Standardizing Patient Outcomes Measurement. *N Engl J Med*, *374*(6), 504-506. https://doi.org/10.1056/NEJMp1511701

Quinten, C., Martinelli, F., Coens, C., Sprangers, M. A., Ringash, J., Gotay, C., Bjordal, K., Greimel, E., Reeve, B. B., Maringwa, J., Ediebah, D. E., Zikos, E., King, M. T., Osoba, D., Taphoorn, M. J., Flechtner, H., Schmucker-Von Koch, J., Weis, J., & Bottomley, A. (2014). A global analysis of multitrial data investigating quality of life and symptoms as prognostic factors for survival in different tumor sites. *Cancer*, *120*(2), 302-311. https://doi.org/10.1002/cncr.28382

Reeve, B. B., Wyrwich, K. W., Wu, A. W., Velikova, G., Terwee, C. B., Snyder, C. F., Schwartz, C., Revicki, D. A., Moinpour, C. M., McLeod, L. D., Lyons, J. C., Lenderking, W. R., Hinds, P. S., Hays, R. D., Greenhalgh, J., Gershon, R., Feeny, D., Fayers, P. M., Cella, D., . . . Butt, Z. (2013). ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Qual Life Res*, *22*(8), 1889-1905. https://doi.org/10.1007/s11136-012-0344-y

Rose, M., Bjorner, J. B., Fischer, F., Anatchkova, M., Gandek, B., Klapp, B. F., & Ware, J. E. (2012). Computerized adaptive testing--ready for ambulatory monitoring? *Psychosom Med*, *74*(4), 338-348. https://doi.org/10.1097/PSY.0b013e3182547392

Rose, M., Bjorner, J. B., Gandek, B., Bruce, B., Fries, J. F., & Ware, J. E., Jr. (2014). The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. *J Clin Epidemiol*, *67*(5), 516-526. https://doi.org/10.1016/j.jclinepi.2013.10.024

Rotenstein, L. S., Huckman, R. S., & Wagle, N. W. (2017). Making Patients and Doctors Happier - The Potential of Patient-Reported Outcomes. *N Engl J Med*, *377*(14), 1309-1312. https://doi.org/10.1056/NEJMp1707537

Russ, T. C., Stamatakis, E., Hamer, M., Starr, J. M., Kivimäki, M., & Batty, G. D. (2012). Association between psychological distress and mortality: individual participant pooled analysis of 10 prospective cohort studies. *BMJ*, *345*, e4933. https://doi.org/10.1136/bmj.e4933

Schalet, B. D., Cook, K. F., Choi, S. W., & Cella, D. (2014). Establishing a common metric for self-reported anxiety: linking the MASQ, PANAS, and GAD-7 to PROMIS Anxiety. *J Anxiety Disord*, *28*(1), 88-96. https://doi.org/10.1016/j.janxdis.2013.11.006

Schalet, B. D., Revicki, D. A., Cook, K. F., Krishnan, E., Fries, J. F., & Cella, D. (2015). Establishing a Common Metric for Physical Function: Linking the HAQ-DI and SF-36 PF Subscale to PROMIS((R)) Physical Function. *J Gen Intern Med*, *30*(10), 1517-1523. https://doi.org/10.1007/s11606-015-3360-0

Schwartz, C. E., Stark, R. B., Borowiec, K., Nolte, S., & Myren, K. J. (2021). Norm-based comparison of the quality-of-life impact of ravulizumab and eculizumab in paroxysmal nocturnal hemoglobinuria. *Orphanet J Rare Dis*, *16*(1), 389. https://doi.org/10.1186/s13023-021-02016-8

Schwarz, R., & Hinz, A. (2001). Reference data for the quality of life questionnaire EORTC QLQ-C30 in the general German population. *Eur J Cancer*, *37*(11), 1345-1351.

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., Gundy, C., Koller, M., Petersen, M. A., Sprangers, M. A., & on behalf of the EORTC Quality of Life Group. (2008). *EORTC QLQ-C30 Reference Values*.

Segawa, E., Schalet, B., & Cella, D. (2020). A comparison of computer adaptive tests (CATs) and short forms in terms of accuracy and number of items administrated using PROMIS profile. *Qual Life Res*, *29*(1), 213-221. https://doi.org/10.1007/s11136-019-02312-8

Sloan, J. A., Loprinzi, C. L., Laurine, J. A., Novotny, P. J., Vargas-Chanes, D., Krook, J. E., O'Connell, M. J., Kugler, J. W., Tirona, M. T., Kardinal, C. G., Wiesenfeld, M., Tschetter, L. K., Hatfield, A. K., & Schaefer, P. L. (2001). A simple stratification factor prognostic for survival in advanced cancer: the good/bad/uncertain index. *J Clin Oncol*, *19*(15), 3539-3546. https://doi.org/10.1200/jco.2001.19.15.3539

Sprangers, M. A., Cull, A., Bjordal, K., Groenvold, M., & Aaronson, N. K. (1993). The European Organization for Research and Treatment of Cancer. Approach to quality of life assessment: guidelines for developing questionnaire modules. EORTC Study Group on Quality of Life. *Qual Life Res*, *2*(4), 287-295. https://doi.org/10.1007/bf00434800

Steppuhn, H., Buda, S., Wienecke, A., Kraywinkel, K., Tolksdorf, K., Haberland, J., Laußmann, D., & Scheidt-Nave, C. (2017). Zeitliche Trends in der Inzidenz und Sterblichkeit respiratorischer Krankheiten von hoher Public-Health-Relevanz in Deutschland. In (Vol. 2): Robert Koch-Institut, Epidemiologie und Gesundheitsberichterstattung.

Stover, A. M., McLeod, L. D., Langer, M. M., Chen, W. H., & Reeve, B. B. (2019). State of the psychometric methods: patient-reported outcome measure development and refinement using item response theory. *J Patient Rep Outcomes*, *3*(1), 50. https://doi.org/10.1186/s41687-019-0130-5

Thom, J., Kuhnert, R., Born, S., & Hapke, U. (2017). 12-month prevalence of self-reported medical diagnoses of depression in Germany. In (Vol. 2): Robert Koch-Institut, Epidemiologie und Gesundheitsberichterstattung.

Valderas, J. M., Ferrer, M., Mendívil, J., Garin, O., Rajmil, L., Herdman, M., & Alonso, J. (2008). Development of EMPRO: a tool for the standardized assessment of patient-reported outcome measures. *Value Health*, *11*(4), 700-708. https://doi.org/10.1111/j.1524-4733.2007.00309.x

Wahl, I., Lowe, B., Bjorner, J. B., Fischer, F., Langs, G., Voderholzer, U., Aita, S. A., Bergemann, N., Brahler, E., & Rose, M. (2014). Standardization of depression measurement: a common metric was developed for 11 self-report depression measures. *J Clin Epidemiol*, *67*(1), 73-86. https://doi.org/10.1016/j.jclinepi.2013.04.019

Wahl, I., Löwe, B., & Rose, M. (2011). Das Patient-Reported Outcomes Measurement Information System (PROMIS®): Übersetzung der Item-Banken für Depressivität und Angst ins Deutsche. *Klinische Diagnostik und Evaluation*, *4*, 236-261.

Waldmann, A., Schubert, D., & Katalinic, A. (2013). Normative data of the EORTC QLQ-C30 for the German population: a population-based survey. *PLoS One*, *8*(9), e74149. https://doi.org/10.1371/journal.pone.0074149

Weinfurt, K. P. (2021). Constructing arguments for the interpretation and use of patient-reported outcome measures in research: an application of modern validity theory. *Qual Life Res*, *30*(6), 1715-1722. https://doi.org/10.1007/s11136-021-02776-7

Weinfurt, K. P. (2022). Constructing and evaluating a validity argument for a performance outcome measure for clinical trials: An example using the Multi-luminance Mobility Test. *Clin Trials*, *19*(2), 184-193. https://doi.org/10.1177/17407745211073609

Wheelwright, S., Bjordal, K., Bottomley, A., Gilbert, A., Martinelli, F., Pe, M., Sztankay, M., Cooks, K., Coens, C., Darlington, A.-S., Fayers, P., Giesinger, J., Koller, M., Kuliś, D., Petersen, M. A., Reijneveld, J., Singer, S., Tomaszewski, K., Fitzsimmons, D., & Group, o. b. o. t. E. Q. o. L. (2021). *EORTC Quality of Life Group guidelines for developing questionnaire modules* (Fifth Edition ed.).

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., & Erikson, P. (2005). Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. *Value in Health*, *8*(2), 94-104.

Acknowledgments

Many people have supported me throughout my academic journey without whom I would not be where I am today. First and foremost I thank my PhD supervisors Professors Richard Osborne and Gerald Elsworth, now at Swinburne University of Technology, Melbourne, Australia, who mentored me throughout my doctoral thesis. I cannot thank you enough for the countless hours we spent discussing my research project and your invaluable feedback that enhanced both my critical thinking and my academic writing. I thank Professor Eckhard Breitbart of Buxtehude, Germany, for the numerous training and networking opportunities you gave me, especially but not limited to the field of skin cancer prevention. Lastly, I thank Professor Matthias Rose at Charité – Department of Psychosomatic Medicine for your trust in letting me reinstate the Health Outcomes Research Unit that I ended up heading for nearly six years. I am grateful for the freedom you gave me to develop the group, grow the team and for letting me develop my own research stream.

Research and teaching is impossible without the many visible but also invisible people that work in the background and ensure that things run smoothly. First of all, I want to thank my former PhD students Gregor Liegl and Kathrin Fischer for your hard work that generated important research output for our group and for all the fun we had both inside and outside of Charité. I also thank the wider Patient-Centered Outcomes Research team at Charité, formerly Health Outcomes Research Unit, for your great collaboration and making the Department of Psychosomatic Medicine a unique place to work. Among others, I especially thank Felix Fischer, Annett Mierke, Alexander Obbarius, Nina Obbarius and Eva Peters. Further, I thank the many people working in the background who helped me navigate the Charité system, particularly Eva Winter, Tobias Hofmann, Agnieszka Seppelt-Górajewska, Ilka Thomalsky-Ritz, Heike Stein, Manuela Hirche, Anja Stielow, the Charité HR, IT and finance teams and many more. Lastly, I thank Christine Kurmeyer and Ingar Abels for your relentless work on supporting women in academia and Professor Ulrike Maschewsky-Schneider for your time and useful tips as part of the Charité mentoring program for women in research.

My thanks further extend to the many formal and especially informal connections that I made during scientific meetings that I attended over the last several years. In particular, I want to acknowledge the wider ISOQOL community that welcomes everyone with open

arms. Ever since attending my first ISOQOL in Lisbon in 2006, I felt supported and was given many opportunities to engage in various roles in ISOQOL's special interest groups, task forces and committees as well as serving on the Board of Directors. In particular, Colleen Pedersen, thank you for your support and getting me involved. Further, I thank several senior colleagues with whom I have engaged over the years during as well as outside of meetings, with some of these informal catch-ups leading to the successful acquisition of research funding to work together. I am grateful to Lori Frank, Theodore Ganiats, Frans Oort, Mirjam Sprangers and Astrid Wahl for your mentorship, particularly during the earlier stages of my career, and Carolyn Schwartz for your ongoing support, advice and friendship. I also thank the wider EORTC community for electing me into the EORTC QLG Executive Committee and giving me the opportunity to participate in the early career investigator program as well as Neil Aaronson, Mogens Grønvold and Morten Petersen for the many thought-provoking conversations we had, especially as part of the EORTC Norm Data project but also during many other occasions. Finally, I want to thank Andrew Bottomley, Director of the EORTC Quality of Life Department, for your ongoing support, and Madeline Pe and Francesca Martinelli of the same department for your support and also for the many fun experiences inside and outside of the EORTC QLG meetings.

Last but not least, I am immensely grateful for the support of my family and friends. Special thanks go to Ursula Nolte, Carsten Nolte, Jennifer Graham and Keith Graham for your love, patience and support in pursuing my goals. Above all, I wholeheartedly thank my husband Duncan Graham and our son Cooper for your ongoing support and your patience when I am spending too much time in front of my laptop, and for teaching me to live in the moment and stop and smell the roses.

Erklärung

§ 4 Abs. 3 (k) der HabOMed der Charité

Hiermit erkläre ich, dass

- weder früher noch gleichzeitig ein Habilitationsverfahren durchgeführt oder angemeldet wurde,
- die vorgelegte Habilitationsschrift ohne fremde Hilfe verfasst, die beschriebenen Ergebnisse selbst gewonnen sowie die verwendeten Hilfsmittel, die Zusammenarbeit mit anderen Wissenschaftlern/Wissenschaftlerinnen und mit technischen Hilfskräften sowie die verwendete Literatur vollständig in der Habilitationsschrift angegeben wurden,
- mir die geltende Habilitationsordnung bekannt ist.

Ich erkläre ferner, dass mir die Satzung der Charité – Universitätsmedizin Berlin zur Sicherung Guter Wissenschaftlicher Praxis bekannt ist und ich mich zur Einhaltung dieser Satzung verpflichte.

…… ……

Datum Unterschrift