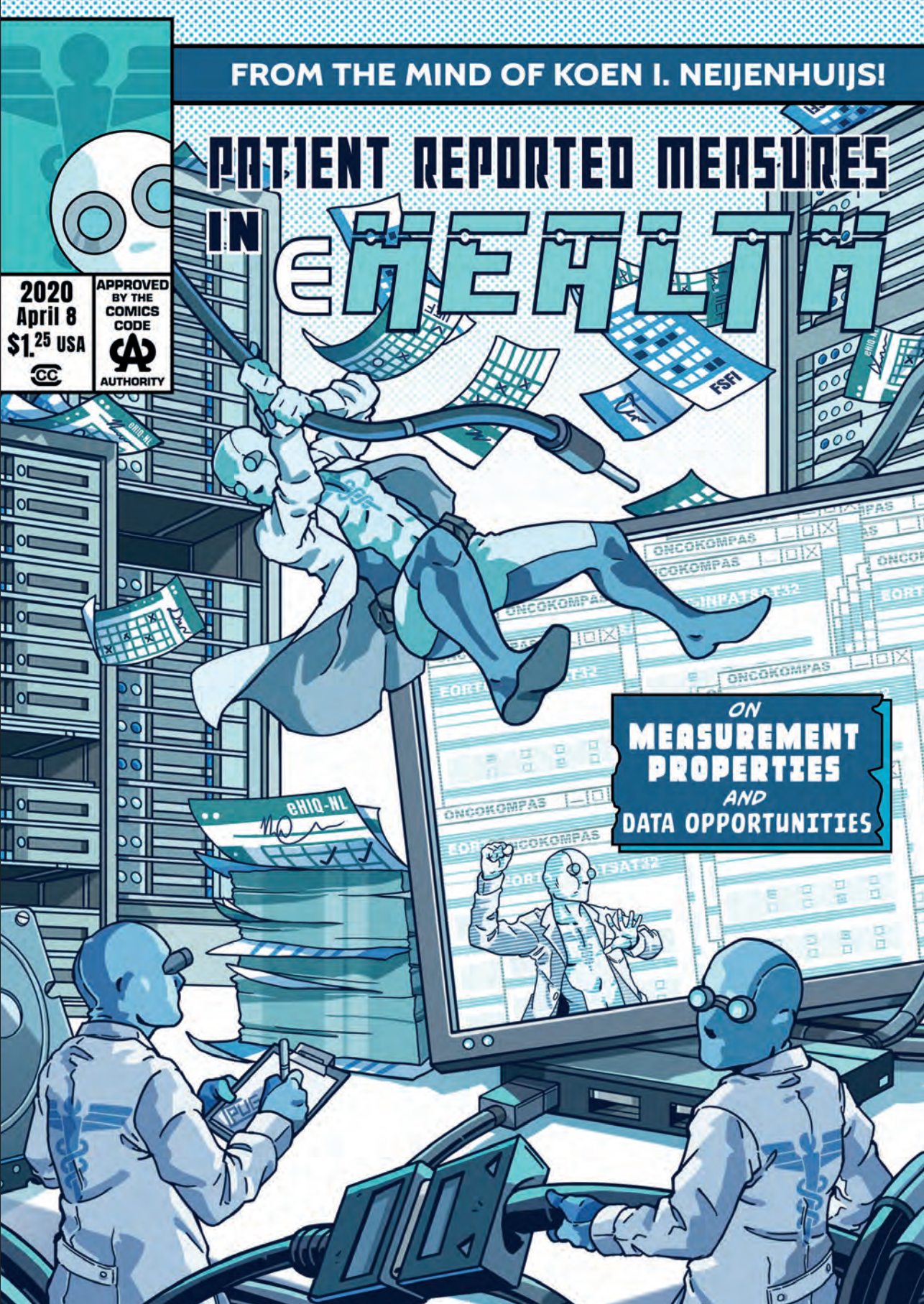
FROM THE MIND OF KOEN I. NEIJENHUIJS!
PATIENT REPORTED MEASURES IN eHEALTH
2020
April 8
$1.25 USA
APPROVED BY THE COMICS CODE AUTHORITY
ON MEASUREMENT PROPERTIES AND DATA OPPORTUNITIES
eHIQ-NL
FSFI
ONCOKOMPAS

Patient reported measures in eHealth: on measurement properties and data opportunities

Koen Ilja Neijenhuijs

ISBN: 978-94-6380-736-4

Cover design by: Chris Puglise || www.artstation.com/cpuglise9

Printed by: ProefschriftMaken || www.proefschriftmaken.nl

VRIJE UNIVERSITEIT

Patient reported measures in eHealth: on measurement properties and data opportunities

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor of Philosophy aan
de Vrije Universiteit Amsterdam
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Gedrags- en Bewegingswetenschappen
op woensdag 8 april 2020 om 11.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Koen Ilja Neijenhuijs
geboren te Deventer

promotoren: prof.dr. I.M. Verdonck - de Leeuw
prof.dr. W.J.M.J. Cuijpers

promotiecommissie: prof.dr. M.M. Riper
prof.dr. M.A.G. Sprangers
prof.dr. W.W. van Solinge
dr. L.B. Mokkink
dr. R.B. Kool

Table of contents

Chapter I

Introduction

To provide adequate health care, the measurement of health and evaluation of health care are important. Measuring health is done in many different ways, ranging from measurement of physical functions (e.g. blood pressure) to an interview between doctor and patient. While physical measurements are a cornerstone of health measurement, many symptoms (e.g. pain, fatigue, mood, anxiety) cannot be measured physically. For such symptoms, we have to rely on a patient's self-report. Even for symptoms that can be measured physically, a patient's self-report is often of additional value. For example, insomnia - and its possible causes - can be measured using polysomnography, which is a combination of multiple physical measurements of body functions during sleep [1]. However, the burden of insomnia on quality of life can only be reported by patients themselves.

A doctor-patient interview provides the advantage of experience and human interpretation. Its disadvantage is that the time that physicians have to interview their patients is often limited [2,3]. To overcome this disadvantage, and due to an increased focus on patient-centred care, the use of patient-reported measures have been promoted by patient organisations, health care providers, and health care insurance companies in the Netherlands [4]. Furthermore, these stakeholders also acknowledge the importance of using patient-reported experience measures to evaluate the quality of health care provision [5].

There are thus two main categories of patient-reported measures (PRMs): Patient-Reported Outcome Measures (PROMs) which aim to measure Health related quality of life (HRQoL) and symptoms of the individual patient, while Patient-Reported Experience Measures (PREMs) aim to evaluate the quality of health care itself from the perspective of the patient. In this dissertation, PROMs and PREMs are central.

I.I Patient-reported outcome measures (PROMs)

Much of the research presented in this dissertation, revolves around the PROMs used in the eHealth application Oncokompas. eHealth is a relatively young and developing field, and pertains to the provision of health care services through digital media [6]. Often eHealth takes the form of a website accessible from computers, phones, and tablets; dedicated software for computers; or dedicated applications for phones and tablets. I will use the term 'eHealth application' as the broad term referring to either eHealth websites, software, or applications. eHealth has been booming in recent years [7], and its use has become widespread throughout the health care trajectory. Oncokompas is an eHealth self-management application that supports Dutch cancer survivors in finding and obtaining optimal supportive care, adjusted to their personal health status and preferences [8–11].

Cancer survivors often experience a wide range of physiological symptoms caused by the disease or by its treatments [12], as well as issues in the psychosocial domain [10,13,14]. The usage of PROMs to monitor HRQoL has been found to be supportive in identifying cancer patients' most bothersome issues [15,16]. While the most bothersome issues differ between individuals, some domains appear to be experienced by many cancer survivors. Some of the most reported issues are: psychological distress such as depression or anxiety [10,13,14,17–19], fatigue [13,14,17,19], pain and pain management [14,17,19], issues stemming from unhealthy lifestyles [10,14,19], role limitations [14,19], problems with cognitive functioning [19,20], sexuality [19,21], and body image [19,22]. Supportive care aims to manage such symptoms and problems and is invaluable in the improvement of HRQoL of cancer survivors [14]. Unfortunately referral rates to relevant supportive care are low [23,24], which was the motivation for the development of Oncokompas [9–11].

Oncokompas entails the components Measure, Learn, and Act. In the Measure component, Oncokompas uses various PROMs to measure HRQoL and symptoms. Within the Measure component, Oncokompas consists of five main quality-of-life domains: physical functioning, psychological functioning, social functioning, lifestyle, and existential issues. Tumour-specific domains are available for patients with breast cancer, colorectal cancer, head and neck cancer, and lymphoma. Each domain is subdivided into subdomains (e.g. sleep issues in physical functioning, depression in psychological functioning). Empirically available cut-off scores and Dutch practice guidelines are used to determine the result for each quality of life domain: "no elevated well-being risk", "elevated well-being risk", or "seriously elevated well-being risk". Based on the results of this Measure component, users are provided with automatically generated, but individually tailored feedback and information on their well-being (Learning), as well as personalized advice on relevant supportive care (Act).

In total, Oncokompas comprises 29 widely used PROMs (besides several other newly developed PROMs). The selection and formulation of all PROMs was performed using a stepwise, iterative, and participatory approach, where non-systematic literature searches were combined with consultations with end-users (i.e. Dutch cancer survivors), health care providers, scientists, and other stakeholders during multiple evaluation cycles. However, the measurement properties of these PROMs were not yet investigated systematically and in detail. Therefore, in this dissertation the measurement properties of the PROMs included in Oncokompas were further investigated.

1.2 Patient-reported experience measures (PREMs)

For the evaluation of health care a priority is put on whether health care is effective. Randomized controlled trials assessing symptom improvement are the norm. However, assessing effectiveness is only half of the story. The way health care is provided can have a large influence on patient outcomes. For example, communication style of primary physicians and their relationship with their patients was found to influence patient adherence to treatment [25,26]. This effect was found to be so profound that health care has shifted to a patient-centred approach [27]. Due to effects such as these, it is important to evaluate the quality of health care provision from the perspective of the patient. PREMs are designed for this specific purpose.

PREMs have been developed since the 1980s, resulting in PREMs that were intended for evaluating health care in general, such as the Patient Satisfaction Questionnaire [28], the Patients' Perceptions of Care Questionnaire [29], the Patients' Consultation Satisfaction Questionnaire [30], the Patient Judgments of Hospital Quality Questionnaire [31], and the Consumer Assessment of Health Plans Study (CAHPS®) 2.0 Adult Core Survey [32]. While certain aspects of quality of care are universal (e.g. communication style of the doctor), many aspects can be very specific to the type of health care. In cancer care, contact with doctors and nurses, as well as extended hospital stays are frequent. To evaluate the specific satisfaction with cancer care, the Quality of Life Group of the European Organisation for Research and Treatment of Cancer (EORTC) developed the IN-PATSAT32 [33]. One aspect that distinguishes the EORTC IN-PATSAT32 from many other PREMs is its international validation, which enables international comparison of patient health care experiences [33].

The evaluation of eHealth applications presents very specific issues. Scientific evaluation using randomized controlled trials and in-depth evaluation through user experience interviews take a lot of time and resources. Meanwhile, the development of eHealth applications is usually rapid, leading to a state of "playing catch-up" for eHealth developers. Furthermore, creating controlled experiments prove difficult to begin with, and confounding variables such as proficiency with the internet can have a large effect on results [34]. While some such standardized measures exist to evaluate usability of software (e.g. the System Usability Scale), they do not offer insights specific to eHealth. To my knowledge, only one PREM has been developed to specifically evaluate eHealth: the eHealth Impact Questionnaire (eHIQ) [35,36]. As such, it is an important tool that requires further investigation. Due to their specific focus and the rigorous methodology used in their development, the EORTC IN-PATSAT32 and the eHIQ are exemplary PREMs. Therefore, in this dissertation the measurement properties of the EORTC IN-PATSAT32 are further investigated; and the eHIQ is translated and validated for the Dutch population of eHealth users.

1.3 Measurement properties

Measurement properties refer to the validity and reliability of a measurement instrument, which are crucial to determine whether the measurement instrument can be used in practice [37]. Validity is "the degree to which a measurement instrument measures the construct(s) it purports to measure", and reliability is "the degree to which the measurement is free from measurement error" [37]. Validity and reliability can be broken down into subcategories (also called measurement properties). The COnsensus-based Standards for the selection of health status Measurement INstruments (COSMIN) taxonomy [37] and COSMIN guidelines [38] provide a framework for discourse and interpretation of these different subcategories, specifically for PRMs. Both the COSMIN taxonomy [37] and COSMIN guidelines [39] were developed in a consensus of 43 experts in in epidemiology, statistics, psychology, and clinical medicine. The COSMIN guidelines were updated in 2018, based on the experience of the use of the COSMIN guidelines in the eight years since its inception [38].

The COSMIN guidelines delineate validity into three subcategories [37]: (i) content validity (the degree to which the content of a PRM is an adequate reflection of the construct to be measured), (ii) construct validity (the degree to which the scores of a PRM are consistent with hypotheses based on the assumption that the instrument validly measures the construct to be measured), and (iii) criterion validity (the degree to which the scores of a PRM are an adequate reflection of a "gold standard", with the gold standard usually being a diagnosis of the symptom to be measured). Construct validity is further delineated into three subcategories: (i) structural validity (the degree to which the scores of a PRM are an adequate reflection of the dimensionality of the construct to be measured), (ii) hypothesis testing (the degree to which the scores of a PRM are consistent with hypotheses based on the assumption that the instrument validly measures the construct to be measured), (iii) cross-cultural validity (the degree to which the performance of the items on a translated or culturally adapted PRM are an adequate reflection of the performance of the items of the original version of the instrument).

Reliability is delineated into three subcategories [37]: (i) internal consistency (the degree of interrelatedness among the items of a PRM), (ii) reliability (the proportion of the total variance in the measurements which is due to "true" differences among patients), and (iii) measurement error (the systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured). Lastly, the COSMIN guidelines define one measurement property outside of the realm of validity and reliability: responsiveness (the ability of a PRM to detect change over time in the construct to be measured).

With each measurement property taking quite a lot of work to test, it can be tempting to pick out the "most important" measurement properties to test and disregard "unimportant" measurement properties. However, each measurement property can be seen as a puzzle piece to be able to determine the applicability and usefulness of a measurement instrument. For example, one might argue that criterion validity is the most important measurement property if you want to use a PROM to approximate a diagnosis of a symptom. But if we have no knowledge of the construct validity - and thus do not know for certain what the instrument measures - can criterion validity truly be interpreted? Investigation of all these measurement properties is of importance for a valid and reliable interpretation of the instrument. The investigation of measurement properties of physical measurements has been rigorous (although criticisms of current practice can be found, e.g. [40]). For example, if we take blood samples to investigate the presence of leukopenia (low white blood cell count), we only accept a small margin of error in the measurement (reliability and measurement error), the cut-off for diagnosis is very clear (criterion validity), and we know that the white blood cell count is directly related to leukopenia (content and construct validity) [41].

1.4 Big data

The use of validated and reliable PRMs in health care creates exciting possibilities. As mentioned, the use of PRMs has been promoted in routine health care in the Netherlands. PRMs are filled in by a patient at various stages of treatment, nowadays often through use of an eHealth application. Through these digitized PRMs an enormous amount of data is gathered. These big data sets can be used to explore theoretical questions that thus far could not be investigated on such a large scale [42]. This data can also be used to develop models able to predict disease trajectories, for example rheumatoid flare-ups [43], cerebral infarction risk [44], and diagnosis of neurological diseases [45]. These large data sets could even be used to investigate the measurement properties of the PRMs themselves, creating an evaluation-loop where PRMs used in health care could be updated and improved over time.

Oncokompas has been used by cancer survivors since 2012 in various research projects as well as in routine care. Hence, a large dataset is currently available including scores of over 1000 Dutch cancer survivors on the 39 (29 pre-existing and 10 newly developed) PROMs, setting a prominent example of data gathered through routine eHealth usage. Symptom clusters are co-occurring symptoms in a group of patients. Symptom clusters have been investigated in cancer patients and survivors, but systematic reviews found little consistency between results of different studies [46,47]. These systematic reviews showed that sample sizes were often too small for the use of appropriate data-analyses, or the amount of symptoms measured was too small. The data from 26 of the 39 PRMs

used in Oncokompas was used to investigate symptom clusters among cancer survivors, using an advanced cluster analysis and network analysis.

Cancer care is inherently complex, with causality of, and interrelatedness between symptoms not always apparent. As mentioned previously, supportive care aims to manage symptoms related to cancer and its treatment, to improve HRQoL [14]; but has a low referral rate [23,24]. Analysis of big datasets can help to unweave the intricate web of causality and interrelatedness of symptoms. In particular, symptoms influencing other symptoms could be identified, which could help with formulating treatment plans targeting first those symptoms that will have the largest impact [48]. Other examples of the usefulness of big datasets are training machine learning algorithms for predicting cancer susceptibility, recurrence, and survival [49,50]. Such algorithms can advise doctors during diagnosis and treatment, and by doing so relieve some of the time burden which limits the time a doctor has for each patient [2,3].

1.5 Aim of this dissertation

The aim of this dissertation is three-fold. The first aim is to investigate the measurement properties of various PROMs included in Oncokompas (chapters 2, and 3). The second aim is to investigate the measurement properties of a widely used PREM in cancer care (chapter 4), and the establishment of a Dutch version of the eHealth Impact Questionnaire (chapter 5). The third aim is to investigate symptom clusters among cancer survivors using a big data set based on PROMs (chapter 6).

In order to investigate the measurement properties of the 29 existing PROMs and one PREM used in Oncokompas, we performed a systematic review. A five-step cascading search strategy was used. First, we searched for systematic reviews of PROMs used in cancer populations. Second, for the PRMs that did not turn up (enough) useable data, we searched for individual validation studies in cancer populations. Third, for the PRMs that did not provide (enough) useable data, we searched for systematic reviews in non-cancer populations. Fourth, for the PRMs that did not turn up (enough) useable data, we searched for individual validation studies in any population. Fifth, for PRMs that had zero hits on the systematic searches, manual searches of the "PROMs in care" database, Google, and Google Scholar were performed. Data was extracted following the COSMIN criteria [37,39]. Data was extracted of 274 studies found in the main systematic searches. For seven PRMs, zero search hits were found, and the manual search resulted in data extraction from six articles, one manual, and one dissertation. Two PRMs had zero usable data sources.

While discussing all of the results of this systematic review is beyond the scope of this dissertation, we delved deeper into the measurement properties of four PROMs and one

PREM that were particularly often-used in practice and research. A report discussing the full results of this systematic review is published elsewhere [51]. In this dissertation, I discuss the measurement properties of two PROMs that aim to assess sexuality. In chapters 2, and 3, we present and discuss the measurement properties International Index of Erectile Function [52,53], and the Female Sexual Function Index [54,55]. The remaining papers on the Body Image Scale [56] and the EORTC QLQ-CR29 [57] are published elsewhere [58,59].

In the second part of this dissertation with a focus on PREMs, the measurement properties of the EORTC IN-PATSAT32 [33] were investigated (chapter 4). After chapter 4, I discuss one of the chance findings of the review in an Intermezzo. Out of 274 articles of which we extracted data on measurement properties, only 13 (<0.05%) reported any information on measurement error. In this Intermezzo we discuss the importance the effect measurement error can have on research and practice, and offer suggestions to improve research into this particular measurement property. In chapter 5 we present the translation and validation of the eHealth Impact Questionnaire; a PREM designed to evaluate eHealth applications from the perspective of its users [35,36].

The third research aim was to answer a research question that could not be reliably investigated without the use of such a unique dataset. In chapter 6 we detail the use of an advanced cluster analysis and network analysis on results from 26 of the PRMs used in Oncokompas to investigate symptom clusters among cancer patients/survivors.

I end the dissertation by discussing implications and future directions of PRMs and big data. I pay particular attention to the effect that insufficiently validated PRMs can have on clinical research and practice, and offer possible solutions to combat these unwanted effects. I also discuss exciting possibilities for using big data, gathered through use of PRMs in eHealth, to improve our health care evaluations and our basic scientific measurements, and to generate and test new hypotheses.


2

Chapter 2

The measurement properties of the IIEF

This chapter was published as Neijenhuijs, K. I., Holtmaat, K., Aaronson, N. K., Holzner, B., Terwee, C. B., Cuijpers, P., & Verdonck-de Leeuw, I. M. (2019). The International Index of Erectile Function (IIEF)—A Systematic Review of Measurement Properties. *The Journal of Sexual Medicine, 16(7),* 1078-1091. doi:10.1016/j.jsxm.2019.04.010.

Abstract

Background: The International Index of Erectile Function (IIEF) is a patient-reported outcome measure to evaluate erectile dysfunction and other sexual problems in males.

Aim: To perform a systematic review of the measurement properties of the IIEF-15 and the IIEF-5.

Methods: A systematic search of scientific literature up to April 2018 was performed. Data were extracted, and analysed according to COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) guidelines for structural validity, internal consistency, reliability, measurement error, hypothesis testing for construct validity and responsiveness. Evidence of measurement properties was categorized into sufficient, insufficient, inconsistent, or indeterminate, and quality of evidence as very high, high, moderate, or low.

Results: Forty studies were included. The evidence for criterion validity (of the Erectile Function subscale), and responsiveness of the IIEF-15 was sufficient (high quality), but inconsistent (moderate quality) for structural validity, internal consistency, construct validity, and test-retest reliability. Evidence for structural validity, test-retest reliability, construct validity, and criterion validity of the IIEF-5 was sufficient (moderate quality), but indeterminate for internal consistency, measurement error and responsiveness.

Clinical Implications: Lack of evidence for and evidence not supporting some of the measurement properties of the IIEF-15 and IIEF-5, shows the importance of further research on the validity of these questionnaires in clinical research and clinical practice.

Strengths & Limitations: A strength of the current review is the use of pre-defined guidelines (COSMIN). A limitation of this review is the use of a precise rather than a sensitive search filter regarding measurement properties to identify studies to be included.

Conclusions: The IIEF requires more research on structural validity (IIEF-15), internal consistency (IIEF-15 and IIEF-5), construct validity (IIEF-15), measurement error (IIEF-15 and IIEF-5), and responsiveness (IIEF-5). The most pressing matter for future research is determining the unidimensionality of the IIEF-5, and the exact factor structure of the IIEF-15.

The International Index of Erectile Function (IIEF) is a widely used patient-reported outcome measure (PROM) to evaluate sexual problems in males [52]. The IIEF is a 15-item PROM (IIEF-15) including five domains: erectile function (6 items), orgasmic function (2 items), sexual desire (2 items), intercourse satisfaction (3 items), and overall satisfaction (2 items). Initial research revealed that the IIEF-15 had acceptable internal consistency ($\alpha > .70$) and test-retest reliability ($r > .70$), except for the orgasmic function scale [52]. Construct validity was good, and the IIEF-15 could detect changes between pre- and post-treatment [52]. A shortened 5-item version was developed to evaluate sexual problems in males by selecting the items that best discriminated between men with and without ED, and adhered to the National Institutes of Heath's definition of ED. The result was a 5-item version consisting of four items from the erectile function, and one item from the sexual intercourse satisfaction subscales. The IIEF-5 was able to discriminate clearly between patients with erectile dysfunction (ED) and those without [54].

Information regarding validity and reliability is of importance for clinical research and practice. To be able to interpret the IIEF-15 and IIEF-5, we need to be certain that the subscales measure what they intend to measure, that they do so consistently, and (particularly for practice) what cut-off scores can be used to screen patients for ED. A review published in 2002 concluded that the IIEF was translated in 32 languages and adopted as a primary endpoint in more than 50 clinical trials worldwide [60]. The authors reported that the IIEF-15 met the standard psychometric criteria for reliability and validity, had a high degree of sensitivity and specificity, and correlated well with other measures of treatment outcome. It also demonstrated good responsiveness [60].

However, since then many more studies have been published investigating the psychometric properties of the IIEF-15 and IIEF-5. Given the high frequency of use in both clinical practice and research, an update of the evidence on the psychometric properties of the IIEF-15 and IIEF-5 is warranted, to investigate whether the initial results [52,54,60] have been replicated in independent international and more recent validation studies. Therefore, the aim of the current study was to perform a systematic review of the measurement properties of the IIEF-15 and IIEF-5.

In this review, we followed the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) methodology [38]. This methodology is based on a taxonomy and definitions of measurement properties for PROMs [39] including content validity, structural validity, internal consistency, cross-cultural validity, reliability, measurement error, criterion validity, hypotheses testing for construct validity, and responsiveness. We hypothesized that there would be evidence supporting sufficient psychometric values IIEF-15 and IIEF-5.

2.1 Methods

2.1.1 Literature search strategy

The literature search was part of a larger systematic review (Prospero ID 42017057237), which investigated the measurement properties of 39 PROMs (including the IIEF-15 and IIEF-5) assessing the quality of life of cancer survivors included in an eHealth application called "Oncokompas" [8–11]. The databases Embase, Medline, and Web of Science were searched using the search terms of the PROM's name and acronyms, combined with a precise filter for measurement properties [61]. The search was performed in January 2017. Appendix A contains the full search terms in regards to all 39 PROMs. Appendix C contains the search terms relating specifically to the IIEF. References were extracted from systematic reviews found in an earlier search of the larger systematic review, and added to the search results. A search update was performed in April 2018. Due to the limitation of the sensitivity of the precise filter (93% sensitive) [61]. a manual search using rudimentary search filters was performed in Google Scholar and Pubmed to check for any prominent records missed in the search update.

2.1.2 Inclusion and exclusion criteria

Studies were included that reported original data on at least one of the following measurement properties of the IIEF as defined by the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) taxonomy [39,62,63]: structural validity (whether the hypothesized measurement model is confirmed), internal consistency (the degree of interrelatedness among the items of the measure), reliability (the proportion of total variance between multiple measurements which is due to "true" differences between measurements), measurement error (a measure of systematic and random error in change scores), criterion validity (whether the measure is an adequate reflection of a gold standard; in the case of the IIEF this is most often a diagnosis of ED), cross-cultural validity (whether the test can be interpreted similarly in different cultures), responsiveness (whether the measure is capable of measuring change over time in the construct to be measured), and hypothesis testing for construct validity (whether the test measures the construct it proposes to measure) which consists of known-groups comparison (a comparison between groups known to have differences on the construct), convergent validity (correlations with other measures that should be related), and divergent validity (correlations with other measures that should be unrelated). While of importance for establishing validity, content validity was not investigated as it was beyond the scope of the current review. Validation studies focused on other PROMs, and non-validation studies that used the IIEF that also reported evidence on the measurement properties of the IIEF were included.

Studies that were only available as abstracts or conference proceedings were excluded, as well as non-English publications. Titles and abstracts, and the selected full-texts were screened by two independent reviewers (KN & MV / KH). Disagreements were discussed until consensus was reached.

2.I.3 Data extraction

Data on each of the measurement properties was extracted by two independent researchers (KN & AvdH / HM / EV / KH). Relevant data included the type of measurement property, its result, and information on methodology. Disagreements were discussed until consensus was reached.

2.I.4 Data analysis

Data analysis was performed in three consecutive steps. First, the methodological quality of the included studies was rated using the 4-point scoring system of the COSMIN checklist [64]. Methodological aspects regarding design requirements and preferred statistical methods specific to each measurement property under consideration, were rated as either "inadequate", "doubtful", "adequate", or "very good". The methodological quality was summarized per measurement property per study as the lowest score received on any of the methodological aspects. Appendix D contains the final study quality ratings.

Second, each measurement property in each individual study was rated as sufficient, insufficient or indeterminate, following the COSMIN guidelines for systematic reviews of PROMs [38]. These ratings were qualitatively summarized to determine the overall rating of the measurement property for the IIEF. If all studies indicated a "sufficient", "insufficient", or "indeterminate" rating for a specific measurement property, the overall rating of this measurement property was rated accordingly. If there were inconsistencies between studies, explanations were explored (e.g. differences in methodological quality, differences in population, etc.). If explanations were found, they were discussed until consensus was reached regarding the overall rating of the measurement property. If no explanations were found, the overall rating would be inconsistent.

Third, the overall rating of evidence per measurement property was supplemented by a level of quality of the evidence, using a modified Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach from the COSMIN methodology [38]. This approach takes into account (i) study quality, (ii) directness of evidence, (iii) inconsistency of results, and (iv) precision of evidence (number of studies and sample size). The overall quality of evidence was rated as high, moderate, low, or very low. Measurement properties that were rated as indeterminate in the previous step did not receive a rating, as there was no evidence to rate.

All ratings (methodological quality, measurement property rating, and GRADE rating) were rated by two independent researchers (KN & KH). Discrepancies in ratings were discussed until consensus was reached.

2.2 Results

2.2.1 Search results

The initial search identified 1401 non-duplicate abstracts of which 568 were relevant to the IIEF (*Figure 2.1*). A total of 526 abstracts and 17 full-texts were excluded as they did not provide unique information on a measurement property. The search update up to April 2018 identified 342 more non-duplicate abstracts. A total of 317 abstracts and 17 full-texts were excluded as they did not provide unique information on a measurement property of the IIEF. A total of 10 references were found through manual means, of which 5 were excluded during abstract screening as they did not provide unique information on a measurement property of the IIEF.

In total, we included forty papers: 31 on the IIEF-15 [52,65–94], 6 on the IIEF-5 [54,95–99], 7 on the IIEF-5 [54,95–100] and 2 on both the IIEF-15 and IIEF-5 [101,102] An overview of study characteristics is provided in Table 2.1. Studies reported sample sizes ranging from 40 to 1764, and 12 different countries were reported: Turkey (Turkish), Spain (Spanish), Taiwan (Taiwanese Mandarin / Hokkien), Germany (German), Iran (Persian), Italy (Italian), Malaysia (Malay), Portugal (Portugese), China (Chinese), Canada (French), Pakistan (Urdu), Netherlands (Dutch). Other included studies likely have been conducted in other countries, but the nationality of participants was not always clearly specified. The combined body of the thirty-three studies on the IIEF-15 and the nine studies on the IIEF-5 reported on all measurement properties, except cross-cultural validity.

Table 2.1. Characteristics of included studies.

Reference	Population	Sample size	Main aim of study
IIEF-15			
Althof et al. (2006) [65]	Patients with ED with somewhat low self-esteem	282	Investigate the impact of sildenafil treatment on pschychosocial functioning and well-being in men with ED from four countries
Bayraktar et al. (2012) [66]	Patients with ED	225	Assess the reliability of the physician-assisted IIEF-15 (Turkish version) in patients with ED
Bayraktar et al. (2013) [67]	Patients with ED	458	To analyze the impact of assistance on the comprehensibility and reliability of the Turkish version of the IIEF-15 questionnaire
Bushmakin et al. (2014) [68]	Patients with ED enrolled in a RCT on sildenafil	500	Testing structural validity of IIEF-15
Cappelleri et al. (1999) [69]	111 ED patients in RCT on sildenafil; 109 control patients; 37 ED patients; and 21 age-matched controls	278	Development and validation of IIEF-15

Reference	Population	Sample size	Main aim of study
IIEF-15			
Cappelleri et al. (2000) [70]	Patients with ED enrolled in a RCT on sildenafil	247	Examine the relationship between patients' self-assessment of EF and the EF domain of the IIEF with respect to ED severity
Cappelleri et al. (2009) [71]	Patients with ED enrolled in a RCT on sildenafil	209	Mapping the relationship between four categories of the EHS and the IIEF-EF, QEQ, SEX-Q, and SEAR
Coyne et al. (2010) [72]	HIV-positive males who have sex with men	486	Validate an adapted version of IIEF-15 for use in HIV-positive males who have sex with men
Flynn et al. (2013) [73]	Cancer patients	389	Validation of the PROMIS sexual function and satisfaction scales
García-Cruz et al. (2011) [74]	Patients referred from general practitioners to urological practice	125	Validate Erection Hardness Score in Spanish
Gelhorn et al. (2017) [75]	Patients diagnosed with hypogonadism	177	Validate the Hypogonadism Impact of Symptoms Questionnaire Short Form
Gonzáles et al. (2013) [76]	Patients participating in a cardiopulmonary or metabolic rehabilitation program	78	Validate the IIEF-15 in Portugese (Brasil) in patients with cardiopulmonary and metabolic diseases
Hwang et al. (2010) [77]	Males aged >30	1060	Assess prevalence of erectile dysfunction in Taiwan
Kriston et al. (2008) [78]	Patients with cardiovascular diseases in rehabilitation centers	261	Test four proposed factor structures of the IIEF-15 in German population
Maasoumi et al. (2017) [79]	Males working in four different work settings	181	Validate the Sexual Quality of Life–Male in Persian (Iran)
Mulhall et al. (2008) [80]	190 men screened for ED ; 902 males participating in a community health survey	1259	Development of Sexual Experience Questionnaire
Nimbi et al. (2018) [81]	Convenience sample	425	Validate the Sexual Modes Questionnaire in Italian
O'leary et al. (2006) [82]	Patients with ED enrolled in a RCT on sildenafil with somewhat low self-esteem	244	Assess the change in confidence, relationship satisfaction and self-esteem in men with ED treated with sildenafil
O'Toole (2018) [83]	Patients with inflammatory bowel disease	175	Develop a IBD-specific Male Sexual Dysfunction Scale
Parisot et al. (2014) [84]	Patients with localized prostate cancer who underwent surgery	75	Validation and responsiveness of Erection Hardness Score
Pascoal et al. (2017) [85]	Heterosexual males in a dyadic relationship	129	Development of the Beliefs About Sexual Functioning Scale
Quek et al. (2002) [86]	20 patients admitted for transurethral resection of the prostate and 20 control males	40	Validate the IIEF-15 in Malaysia
Quinta Gomes et al. (2012) [87]	Sexually healthy males and patients with ED	1363	Validate the IIEF-15 in Portugal
Rosen et al. (1997) [52]	111 patients with ED part of a sildenafil RCT; 109 matched healthy men; 37 patients with ED; 21 matched healthy controls	278	Development and first validation of IIEF-15
Rosen et al. (2011) [88]	Participants in RCT on tadalafil	863	Estimate Minimal Clinically Important Difference for the Erectile Function subscale of the IIEF-15

Reference	Population	Sample size	Main aim of study
IIEF-15			
Rubio-Aurioles et al. (2009) [[89]]	51 couples with untreated ED; 57 couples without ED	107	Development and first validation of the Female Assessment of Male Erectile
Saffari et al. (2016) [90]	Males attending a health post	1764	Validate the Male Genital Self-Image Scale for Iranian Men
Serefoglu et al. (2008) [91]	Patients from an urology clinic	430	Analyze the impact of patient age, education level, and household income on the comprehension of the IIEF-15 (Turkish version) and determine the patient characteristics that make this questionnaire less reliable
Tang et al. (2018) [92]	260 patients diagnosed with premature ejaculation, and 104 healthy controls	364	Validate the Premature Ejaculation Diagnostic Tool in Chinese
Terrier et al. (2017) [93]	Sexually active patients with early-stage prostate cancer after radical prostatectomy	178	Define the optimal Erectile Functioning score that optimally defines "functional" erections after radical prostatectomy
Wiltink et al. (2003) [94]	59 ED patients, 38 patients with Peyronie's disease, and 33 control males	130	Validate IIEF-15 for the German population (Germany)
IIEF-15 & IIEF-5			
Dargis et al. (2013) [101]	Canadian males aged > 65 years	508	Validation of IIEF-15 and IIEF-5 in an older population
Lim et al. (2003) [102]	111 healthy males; 60 patients attending primary care clinics; 32 ED patients undergoing sildenafil therapy	197	Validate the IIEF-15 and IIEF-5 in Malay (Malaysia)
IIEF-5			
Aslan et al. (2011) [95]	Patients with ED	81	Evaluate the association between IIEF-5 and Erection Hardness Grade Score in patients who underwent sildenafil citrate treatment for ED
Cappelleri et al. (2001) [100]	Patients with ED enrolled in a RCT on sildenafil	247	Examine the relationship between patients' self-assessment of EF and classification of ED severity using the IIEF-5
Lin et al. (2016) [96]	Prostate cancer patients in sexual relationships	1058	Rasch analysis of Premature Ejaculation Diagnostic Tool and IIEF-5 in Iranian prostate cancer patients
Mahmood et al. (2012) [97]	Patients from an urology clinic	47	Validate the IIEF-5 in Urdu (Pakistan)
Rosen et al. (1999) [53]	1063 patients with ED enrolled in a sildenafil RCT, and 116 healthy controls	1152	Development of an abridged version of the IIEF-15 (the IIEF-5)
Tang et al. (2015) [98]	Patients diagnosed with LPE, heterosexual with a sexual relationship of over 6 months	406	Validate IIEF-5 for erectile function in Lifelong Premature Ejaculation patients in China
Utomo et al. (2015) [99]	82 ED patients; 253 controls	335	Validate IIEF-5 in Dutch (Netherlands)
Utomo et al. (2015) [99]	82 ED patients; 253 controls	335	Validate IIEF-5 in Dutch (Netherlands)

IIEF: International Index of Erectile Function; ED: Erectile Dysfunction; EF: Erectile Function; EHS: Erection Hardness Score; QEQ: Quality of Erection Questionnaire; SEX-Q: Sexual Experience Questionnaire; SEAR: Self-Esteem And Relationship questionnaire; RCT: Random Controlled Trial; PROMIS: Patient-Reported Outcomes Measurement Information System; IBD: Inflammatory Bowel Disease; LPE: Lifelong Premature Ejaculation

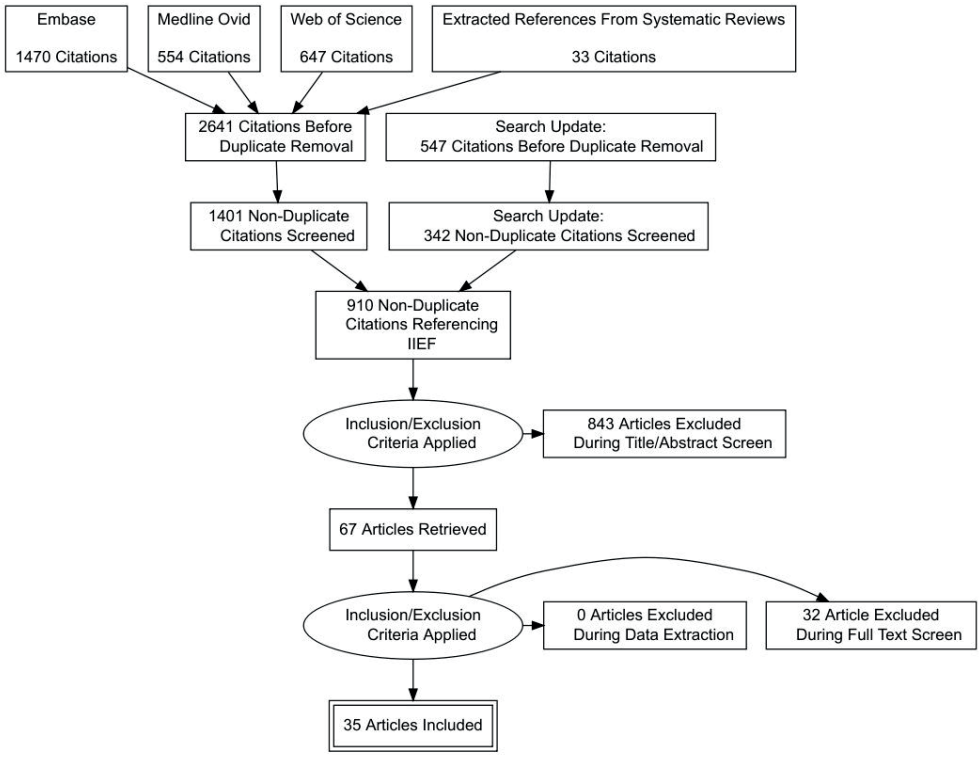

Figure 2.1. PRISMA diagram.

2.2.2 Structural validity

Eight studies reported on structural validity of the IIEF-15 [52,68,72,76,78,87,94,102], of which one study [87] reported two types of analyses (*Table 2.2*). Methodological quality was rated as "very good" [68,78], "adequate" [52,72,94,102], or "doubtful" [76,87]. One "doubtful" score was due to an insufficient sample size ("other flaws" in COSMIN methodological quality) [76], while the other was for very unequal subgroup sizes ("other flaws" in COSMIN methodological quality) [87].

Three studies of "very good" [68,78], and "doubtful" [87] quality, reported Confirmatory Factor Analyses (CFAs). The evidence on structural validity was rated as sufficient in two studies, as a good fit was found for a 5-factor structure [78,87]. The evidence was rated as insufficient for the third study, as the fit for the 5-factor structure was below acceptable levels (Comparative Fit Index [CFI] < .95) [68]. The evidence was rated as indeterminate for six studies of the IIEF-15, of "adequate" [52,72,94,102] and "doubtful" [76,87] quality, as they reported Principal Component Analyses (PCAs) without fit measures.

Notably, two of these studies reproduced the hypothesized five components, two studies found four components, and two studies found two components.

One study reported on structural validity of the IIEF-5 [96] (*Table 2.2*). Methodological quality was rated as "very good". Evidence on structural validity was rated as sufficient, as a good fit of a Rasch model was reported.

Table 2.2. Structural validity of the IIEF.

Reference	Methodology	Outcome	Rating	Quality
IIEF-15				
Bushmakin et al. (2014) [68]	Confirmatory Factor Analysis	5-factor solution found on baseline (N=500; CFI=.92); on end of DBPC phase (N=458; CFI=.94); and end of open-label (N=454; CFI: .93), all with bad fit (CFI < .95).	Insufficient	Very good
Coyne et al. (2010) [72]	Principal Component Analysis	Four factors with Eigenvalue > 1.5. The original domains of intercourse and overall satisfaction appeared together in one factor.	Indeterminate	Adequate
Gonzáles et al. (2013) [76]	Principal Component Analysis	Five factors explaining 75.8% of variance; most questions were loaded correctly on their respective domains, except for sexual satisfaction domain, which comprises questions 6, 7, and 8, which presented a confounding factor. Question 1 equally loaded on two factors.	Indeterminate	Doubtful*
Kriston et al. (2008) [78]	Confirmatory Factor Analysis	Original five factor model had acceptable fit (GFI = .889; TLI = .933; CFI = .949; SRMR = .045; RMSEA = .09) as did a four-factor model (GFI = .849; TLI = .908; CFI = .926; SRMR = .049; RMSEA = .107). A two-factor model had non-acceptable fit (CFI = .783; TLI = .854; CFI = .876; SRMR = .064; RMSEA = .134), as did a one-factor model (GFI = .743; TLI = .812; CFI = .839; SRMR = .072; RMSEA = .152). CAIC favored the original five factor model (512.68).	Sufficient	Very good
Lim et al. (2003) [102]	Principal Component Analysis	The expected structure of five distinct domains was not clearly present. The eigenvalue was concentrated on the first factor, while the remaining four factors extracted had eigenvalue less than 1. Factor 2 of the Malay version of IIEF corresponded with the OS domain of the original IIEF, while factor 3 corresponded with SD domain, and Factor 4 with OF domain. Factor 1 contained a mixture of loadings from both EF and IS domains.	Indeterminate	Adequate

Reference	Methodology	Outcome	Rating	Quality
IIEF-15				
Quinta Gomes et al. (2012) [87]	Principal Component Analysis	2 components explaining 55% variance. The first component cluster loadings from eight items of the erection and orgasm domains of the original IIEF. The second component included the original dimensions of SD, IS, and OS, was composed of the remaining six items of the scale.	Indeterminate	Doubtful **
Quinta Gomes et al. (2012) [87]	Confirmatory Factor Analysis	Acceptable fit for 2-factor model (RMSEA = .077; CFI = .94; GFI = .93; AGFI = .90) and 5-factor model (RMSERA = .067; CFI = .96; GFI = .95; AGFI = .92)	Sufficient	Doubtful
Rosen et al. (1997) [52]	Principal Component Analysis	Five factor solution. (1) erectile function, (2) orgasmic function, (3) sexual desire, (4) intercourse satisfaction, and (5) overall satisfaction.	Indeterminate	Adequate
Wiltink et al. (2003) [94]	Principal Component Analysis	Two factors found explaining 70% variance. First factor (12 items) of sexual function. Second factor (3 items) of sexual desire.	Indeterminate	Adequate
IIEF-5				
Lin et al. (2016) [96]	Rasch analysis	Monotonical increase across IIEF; one local dependency in IIEF; no substantial DIF in IIEF	Sufficient	Very good

IIEF: International Index of Erectile Function; EF: Erectile Function; OF: Orgasmic Function; SD: Sexual Desire; IS: Intercourse Satisfaction; OS: Overall Satisfaction; GFI: Goodness of Fit Index; TLI: Tucker Lewis index; CFI: Comparative Fit Index; SRMR: Standardised Root Mean Square Residual; RMSEA: Root Mean Square Error of Approximation * Due to insufficient sample size ** Due to very unequal subgroup sizes

2.2.3 Internal consistency

Fifteen studies reported on internal consistency of the IIEF-15 [52,66,67,72,76,78,81,85–87,89,92,94,101,102] (*Supplementary Table 7.1*). Methodological quality of these studies was rated as "very good" [52,67,72,78,87,89,101,102], "adequate" [76,81,94], or "inadequate" [67,85,86,92]. The inadequate scores were due to only reporting internal consistency for the total IIEF-15 instead of its' subscales [67,85,92] or because of a very small sample size ("other flaws" in COSMIN methodological quality) [86].

Eight studies, of "very good" [66,78,87,101], "adequate" [81], and "inadequate" [85,86,92] quality, reported Cronbach's Alpha of sufficient values of the IIEF-15. Five studies, of "very good" [52,72,89,102], and "adequate" [76] quality, reported Cronbach's Alpha of insufficient values of the IIEF-15. In two studies the evidence on internal consistency was rated as indeterminate as it could not be interpreted: one study did not report the internal consistency per subscale [67], and one study reported internal consistency for two subscales resulting from their PCA results [94].

Five studies reported on internal consistency of the IIEF-5 [97–99,101,102] (*Supplementary Table 7.1*). Methodological quality of these studies was rated as "very good" [98,99,101,102], or "inadequate" [97]. The inadequate score was due to a very small N ("other flaws" in COSMIN methodological quality) [97]. The evidence of internal consistency was rated as indeterminate for all five studies, as unidimensionality was not investigated (see Structural Validity), which is a prerequisite for internal consistency.

2.2.4 Test-retest reliability

Eight studies reported on test-retest reliability of the IIEF-15 [52,66,67,76,86,87,91,102] (*Table 2.3*). Methodological quality of these studies was rated as "doubtful" [52,76,87,91,102], or "inadequate" [66,86]. The doubtful scores were due to inappropriate time intervals (the same day) [91,102], and reporting of correlation coefficients instead of the Intraclass Correlation Coefficient [52,67,76,87,91]. The inadequate scores were due to test conditions that differed across measurements [66], and a very small N ("other flaws" in COSMIN methodological quality) [86].

The evidence on test-retest reliability was rated as sufficient in five studies, of "doubtful" [52,76,102], and "inadequate" [66,86] quality. The evidence was rated as insufficient in two studies, of "doubtful" [87,91] quality, because reported values of reliability were below .70. The evidence was rated as indeterminate in one study, of "doubtful" [67] quality, as the values were subdivided in six subgroups and not well interpretable.

Two studies reported on test-retest reliability of the IIEF-5 [99,102]. Methodological quality was rated as "adequate" [99], or "doubtful" [102]. The doubtful score was due to inappropriate time intervals (the same day) [102]. The evidence on test-retest reliability in both studies was rated as sufficient.

Table 2.3. Test-retest reliability of the IIEF.

Reference	Coefficient	IIEF5	Total score	EF	OF	SD	IS	OS	Rating	Quality
IIEF-15										
Bayraktar et al. (2012) [66]	Correlation		0.91	0.94	0.83	0.87	0.75	0.78	Sufficient	Inadequate***
Bayraktar et al. (2013) [67]	Rho		.39 - .87						Indeterminate	Doubtful**
Gonzáles et al. (2013) [76]	ICC		.80 - .98	.90 - .98	.91 - .98	.80 - .92	.82 - .97	.89 - .98	Sufficient	Doubtful

Reference	Coefficient	IIEF-5	Total score	EF	OF	SD	IS	OS	Rating	Quality
IIEF-15										
Quek et al. (2002) [86]	ICC			0.77	0.75	0.87	0.79	0.85	Sufficient	Inadequate****
Quinta Gomes et al. (2012) [87]	Correlation			0.55	0.69	0.14	0.71	0.9	Insufficient	Doubtful**
Rosen et al. (1997) [52]	Correlation		0.82	0.84	0.64	0.71	0.81	0.77	Sufficient	Doubtful**
Serefoglu et al. (2008) [91]	Kappa		0.37						Insufficient	Doubtful***
IIEF-15 & IIEF-5										
Lim et al. (2003) [102]	ICC	0.88	0.92	0.88	0.82	0.82	0.89	0.82	Sufficient	Doubtful*
IIEF-5										
Utomo et al. (2015) [99]	ICC	0.88							Sufficient	Adequate

IIEF: International Index of Erectile Function; EF: Erectile Function; OF: Orgasmic Function; SD: Sexual Desire; IS: Intercourse Satisfaction; OS: Overall Satisfaction * Due to inappropriate time intervals ** Due to reporting of inappropriate coefficients *** Due to test conditions differing across measurements **** Due to an extremely small N

2.2.5 Measurement error

One study reported measurement error of IIEF-15 [86], and measurement error was calculated for one study which reported test-retest reliability [52] (*Supplementary Table 7.2*). Methodological quality was rated as "adequate" [52] or as "inadequate" [86]. The inadequate rating was due to a very small N ("other flaws" in COSMIN methodological quality) [86].

For interpretation of measurement error, the Minimal Clinically Important Difference (MCID) is necessary. The evidence on measurement error was rated as indeterminate for the two studies [52,86] as no MCID was reported for any of the subscales in any of the included studies, except for the Erectile Function subscale for which a MCID was reported (mean MCID = 7.27) [88].

The evidence on measurement error of the Erectile Function subscale was rated as insufficient for one study [86], for which we could calculate the SEM (0.69 - 3.59) and the Smallest Detectable Change (SDC; 1.90 - 9.94). The SDC is the minimum change score necessary to have 95% confidence that it represents a true change. The MCID is the smallest change score that represents a clinically relevant change. The SDC should be smaller than the MCID, so that a smallest clinically relevant change score can be

distinguished from measurement error. In this case, the SDC (9.49) was larger than the MCID (7.27), leading to an insufficient rating for the Erectile Function subscale.

One study reported measurement error of the IIEF-5 [99]. Methodological quality was rated as "adequate". Limits of Agreement (LoA) were reported (10.1). Evidence on measurement error was rated as indeterminate, as no MCID or MIC was reported.

2.2.6 Construct validity (hypothesis testing)

2.2.6.1 Known-group comparison

Seven studies reported known-group comparison of the IIEF-15 [52, 86, 87, 92, 94, 101, 102] (*Supplementary Table 7.3*). Known group differences were investigated in relation to age [101], diagnosis of ED [52, 87, 94, 102], diagnosis of premature ejaculation [92], lifelong versus acquired premature ejaculation [92], and treatment versus control [86]. The methodological quality was rated as "adequate" [52,87,92,94,101,102] or "inadequate" [86]. The inadequate rating was due to a very small N ("other flaws" in COSMIN methodological quality) [86]. Evidence for construct validity was rated as sufficient for all studies.

Two studies reported known-group comparison of the IIEF-5 [54,101], and compared age groups [101], and diagnosis of ED [54]. The methodological quality was rated as "adequate" [101] or "doubtful" [54]. The doubtful rating was due to very unequal group sizes ("other flaws" in COSMIN methodological quality) [54]. Evidence of construct validity was rated as sufficient.

2.2.6.2 Convergent validity

Seventeen studies reported on convergent validity of the IIEF-15 [52,70,71,73–75,77,79–81,83–85,89,90,92,94] (*Supplementary Table 7.4*). The IIEF-15 was compared to a single item self-assessment of ED [70], the PROMIS sexual domain (Patient Reported Outcomes Measurement Information System [73]), Quality Erection Questionnaire [77], Erection Hardness Score [71,74,77,84], Sexual Experience Questionnaire [80], Male Genital Self-Image Scale [90], Female Assessment of Male Erection [89], partnership satisfaction [94], Hypogonadism Impact of Symptoms Questionnaire Short Form [75], Sexual Quality of Life–Male [79], Sexual Modes Questionnaire [81], Inflammatory Bowel Disease Male Sexual Dysfunction Scale [83], Beliefs About Sexual Functioning Scale [85], Premature Ejaculation Tool [92], and clinician ratings [52,89,94].

The methodological quality was rated as "adequate" [52,73,77,79,81,83,85,89,90,92,94] or "doubtful" [70,71,74,75,84]. The doubtful ratings were due to a small N ("other flaws" in COSMIN methodological quality) [84], use of Pearson correlation where

Spearman correlation should have been used [74], imprecise reporting of hypotheses ("other flaws" in COSMIN methodological quality) [75], the lack of information on measurement properties of the comparator instrument [70], or imprecise reporting of results [71].

The evidence on construct validity was rated as sufficient for eleven studies, of "adequate" [52,73,77,79,80,89,94] and "doubtful" [70,74,75,84] quality. The evidence was rated as insufficient for five studies, of "adequate" [81,83,85,90,92] and one study of "doubtful" [71] quality, as reported correlations were low.

Two studies reported on convergent validity of the IIEF-5 [95,100], and compared the IIEF-5 to the Erection Hardness Scale [95], a single item self-assessment of ED [100], the Erectile Dysfunction Inventory of Treatment Satisfaction [100], 5-item version of the Erectile Dysfunction Inventory of Treatment Satisfaction filled in by a partner [100], and a single item of global efficacy of erections [100]. Methodological quality was rated as "adequate" [95] or "doubtful" [100]. The doubtful rating was due the lack of information on measurement properties of the comparator instrument [100]. The evidence on construct validity was rated as sufficient for one study [95], and insufficient for one study [100], as the reported correlation was low.

2.2.6.3 Divergent validity

Three studies reported on divergent validity of the IIEF-15 [52,94,101] (*Supplementary Table 7.5*), and compared the IIEF-15 to the Dyadic Adjustment Test and SF-12 [101], the Locke-Wallace Marital Adjustment Test [52], State-Trait Anxiety Inventory, Center for Epidemiological Studies Depression Scale [94] and social desirability [52,94]. Methodological quality was rated as "adequate" [94,101] or "doubtful" [52]. The doubtful score was due to non-reporting of measurement properties of the comparison instrument. The evidence on construct validity was rated as sufficient for all studies.

One study reported on divergent validity of the IIEF-5 [101] (*Supplementary Table 7.5*), and compared the IIEF-5 to the Dyadic Adjustment Test and SF-12. Methodological quality was rated as "adequate", and evidence was rated as sufficient.

2.2.7 Criterion validity

Four studies reported on criterion validity of the IIEF-15 Erectile Function subscale [69,89,93,94] (*Table 2.4*). One study also reported criterion validity for the IIEF-15 total score [94]. Methodological quality was "very good" [69,89], "adequate" [94], or "doubtful" [93]. The "doubtful" rating was due to use of a questionable gold standard (intercourse satisfaction). All other studies used ED diagnosis as the gold standard.

The evidence on criterion validity was rated as sufficient for three studies, of "very good" [69,89] and "doubtful" [93] quality. Two studies [69,89] reported Area Under the Curve (AUC) values for the Erectile Function subscale as .97 for diagnosing ED, with good sensitivity (.97 - .98) and specificity (.79 - .88) for the cut-off point of 25. One study [93] reported an AUC value for the Erectile Function subscale as .86 for determining intercourse satisfaction. Good sensitivity (.77 and .78) and specificity (.92 and .80) were reported for the cut-off points of 24 and 25, respectively. The evidence was rated as indeterminate for one study [94], as no AUC value was reported.

Three studies reported on criterion validity of the IIEF-5 [54,98,102] (*Table 2.4*). Methodological quality was "very good" [98], "adequate" [102], or "doubtful" [54]. The doubtful rating was due to very unequal group sizes [54]. The evidence on criterion validity was rated as sufficient for all studies, with reported AUC between .86 - .97 [54,98,102]. All studies reported good sensitivity (.85 - .98) and specificity (.75 - .88) for cut-off points of 15.5, 17, and 21.

Table 2.4 Criterion validity of the IIEF.

Reference	Instrument	AUC	Cut.off	Sensitivity	Specificity	PPV	NPV	Rating	Quality
IIEF-15									
Cappelleri et al. (1999) [69]	IIEF-15 EF	0.97	25	0.97	0.88	0.89	0.97	Sufficient	Very good
Rubio-Aurioles et al. (2009) [89]	IIEF-15 EF	0.97	25	0.98	0.79			Sufficient	Very good
Terrier et al. (2017) [93]	IIEF-15 EF	0.86	24 25	.78 .77	.80 .82			Sufficient	Doubtful*
Wiltink et al. (2003) [94]	IIEF-15 Total		53	0.87	0.75	0.85		Indeterminate	Adequate
	IIEF-15 EF		21	0.84	0.72	0.84			
IIEF-5									
Lim et al. (2003) [102]	IIEF-5	0.86	17	0.85	0.75			Sufficient	Adequate
Rosen et al. (1999) [53]	IIEF-5	0.97	21	0.98	0.88	0.89	0.98	Sufficient	Doubtful**
Tang et al. (2015) [98]	IIEF-5	0.97	22	1	0.06			Sufficient	Very good
			15.5	0.97	0.86				

IIEF: International Index of Erectile Function; AUC: Area Under the Curve; PPV: Positive Predictive Value; NPV: Negative Predictive Value; CART: Classification and Regression Trees * Due to a doubtful criterion ** Due to very unequal group sizes which biases the results of the CART algorithm; and due to usage of training sample in cross-validation

2.2.8 Responsiveness

Six studies reported responsiveness of the IIEF-15 [52,65,70,82,84,86] (*Supplementary Table 7.6*). Methodological quality was rated as "adequate" [52,65,70,82,84], or "inadequate" [86]. The inadequate rating was due to a very small N ("other flaws" in COSMIN methodological quality) [86]. The evidence on responsiveness was rated as sufficient for all six studies.

Two studies reported on responsiveness of the IIEF-5 [99,100] (*Supplementary Table 7.6*). Methodological quality was rated as "adequate" [100] or "doubtful" [99]. The doubtful rating was due to a very small group of treated patients ("other flaws" in COSMIN methodological quality). The evidence on responsiveness was rated as sufficient for both studies.

2.2.9 Data synthesis

The overall ratings of the measurement properties can be found in Table 2.5.

Table 2.5. Ratings of measurement properties.

Measurement Property	Rating of Measurement Property	Quality of Evidence
IIEF-15		
Structural Validity	Inconsistent	Moderate
Internal Consistency<U+2060>	Inconsistent	Moderate
Reliability	Inconsistent	Moderate
Measurement Error	Indeterminate / Insufficient (Erectile Function subscale)	Very low
Construct Validity	Inconsistent	Moderate
Criterion Validity	Sufficient	High
Responsiveness	Sufficient	High
IIEF-5		
Structural Validity	Sufficient	Moderate
Internal Consistency<U+2060>	Indeterminate	
Reliability	Sufficient	Moderate
Measurement Error	Indeterminate	
Construct Validity	Sufficient	High
Criterion Validity	Sufficient	Moderate
Responsiveness	Indeterminate	

Structural validity of the IIEF-15 was rated as inconsistent with evidence of moderate quality, due to the inconsistencies in the findings. Structural validity of the IIEF-5 was rated as sufficient with evidence of moderate quality, as it was based on only one study.

Internal consistency of the IIEF-15 was rated as inconsistent with evidence of moderate quality, due to inconsistencies in the findings. Internal consistency of the IIEF-5 was rated as indeterminate, due to the lack of evidence for unidimensionality.

Reliability of the IIEF-15 was rated as inconsistent with evidence of moderate quality, due to inconsistencies in the findings. Reliability of the IIEF-5 was rated as sufficient with evidence of moderate quality, due to some risk of bias resulting from the methodological quality. For both IIEF-15 and IIEF-5, measurement error was rated indeterminate, except for the erectile function scale which was rated as insufficient.

Construct validity (hypothesis testing) of the IIEF-15 was rated as inconsistent with evidence of moderate quality. Eleven studies showed sufficient scores, while six studies showed insufficient scores. We note that some of the comparator instruments in convergent validity are of questionable relevance (e.g. the Male Genital Self-Image Scale) or quality (e.g. comparators that were only validated once in their lifetime). As such, while formally rating the construct validity of the IIEF-15 as inconsistent, the rating leans more to sufficient than insufficient. Construct validity of the IIEF-5 was rated as sufficient with evidence of high quality. One study showed values of insufficient convergent validity of the IIEF-5, these values were only just below sufficient levels, and were discounted against the evidence for sufficient construct validity.

Criterion validity was rated as sufficient and evidence of high quality for the IIEF-15, and evidence of moderate quality for the IIEF-5 due to some risk of bias resulting from the methodological quality. Responsiveness was rated as sufficient and evidence of high evidence for the IIEF-15, and as indeterminate for the IIEF-5.

2.3 Discussion

This systematic review investigated the evidence regarding the measurement properties of the IIEF-15 [52] and IIEF-5 [54]. In contrast to our hypothesis, the majority of the measurement properties were not rated as sufficient for both the IIEF-5 and IIEF-15. The IIEF-15 was rated as sufficient on criterion validity (of the Erectile Function subscale), and responsiveness, with sufficient ratings with high level of evidence. The evidence for structural validity, internal consistency, construct validity, and test-retest reliability were rated inconsistent, with moderate level of evidence. Measurement error

for the Erectile Function subscale was rated as insufficient with very low quality of evidence, while it was indeterminate for the remaining subscales.

The IIEF-5 was rated as sufficient on criterion validity with high quality of evidence. The IIEF-5 was also rated as sufficient on structural validity, test-retest reliability, and construct validity, but with moderate quality of evidence as the evidence was based on very few studies. The evidence for internal consistency, measurement error and responsiveness were rated as indeterminate.

With regard to structural validity, there is some evidence from CFAs [78,87] and PCAs [52,76] that the IIEF-15 consists of a 5-factor structure as hypothesized [52]. However, there is also evidence not supporting the 5-factor structure: one CFA found a poor fit for a 5-factor structure [68], one CFA found acceptable fits for both a 2-factor (one factor of erectile function and orgasm, and one factor of desire and satisfaction) and 5-factor structure [87], one CFA found acceptable fits both a 4-factor (combined factor of erectile function and intercourse satisfaction) and 5-factor structure [78] and multiple PCAs found either a 4-factor solution (combined component of erectile function and intercourse satisfaction [102], or combined component of intercourse satisfaction and overall satisfaction [72]), or a 2-factor solution (one component of erectile function and orgasm, and one component of desire and satisfaction [87], or one component of sexual function and one component of sexual desire [94]). There seems to be as much, if not more evidence against the 5-factor structure.

The results of the current review are in line with the concerns raised by Forbes et al. [103,104] that the five-factor structure is not as firmly established as argued by Rosen et al. [60,105]. We agree with Rosen et al.'s reply [105] that low correlations between subscales of the IIEF-15 do not warrant an insufficient rating of structural validity, but disagree with their underrating for the concerns regarding the structural validity of the IIEF-15. Their evidence cited concerns exploratory factor analyses, with no mention of confirmatory analyses which provide a higher level of evidence for structural validity. Two of the confirmatory analyses we identified showed evidence for both the five-factor structure and alternative factor structures [78,87], and the remaining CFA showed evidence against the five-factor structure [68]. Future studies are clearly needed to investigate alternative factor structures (e.g. 2-factor, 4-factor, second-order hierarchical factors) and compare them directly to the posited 5-factor structure.

The structural validity of the IIEF-5 is also of interest. While one Rasch analysis showed sufficient structural validity, no tests of unidimensionality were reported in any of the included articles. The IIEF-5 consists of items representing both erectile dysfunction

(items 2, 4, 5, and 15 from the IIEF-15), as well as sexual intercourse satisfaction (item 7 from the IIEF-15). Theoretically, the IIEF-5 may be multidimensional due to the use of two constructs during development. Tests of unidimensionality are of importance to further determine the structural validity of the IIEF-5.

The internal consistency of the IIEF-15 showed values that were very high indicating possible redundancy (Alpha > .95; 3 studies of very good quality), as well as values considered too low (Alpha < .70; 1 study of very good quality). However, many studies (12 studies of inadequate to very good quality) showed sufficient internal consistency. The methodological quality is of importance to put these values in context, where an equal number of very good quality studies found insufficient as sufficient values. Considering these results, it is possible that internal consistency of the IIEF-15 may vary across subgroups. However, when examining the populations of the studies that reported sufficient values [67,78,81,85–87,92,101] versus those of the studies that reported insufficient values [52,72,76,89,102] no clear pattern arose, with both groups of studies investigating different nationalities as well as subgroups (e.g. older men, HIV-positive men who have sex with men, sexually healthy men, men suffering from ED). Furthermore, these inconsistencies may be caused by differences in factor structure across subgroups. A future cross-cultural study design, investigating measurement invariance, may help elucidate the inconsistencies of these findings.

The evidence on internal consistency of the IIEF-5 can not yet me determined, as the unidimensionality (a prerequisite for internal consistency) has not yet been tested. However, if unidimensionality is tested and found to be sufficient, internal consistency is likely to be rated as sufficient. One study (of very good quality) found an insufficient value (Alpha < .70) while 3 studies of very good quality found sufficient values.

While both the IIEF-15 Erectile Function subscale and the IIEF-5 showed to be able to sufficiently predict ED diagnosis, it is not yet clear which cut-off scores are most suitable. Making a direct comparison between sensitivity and specificity ratings of cut-off scores across studies is beyond the score of the current review, as an individual patient meta-analysis would be required. Furthermore, a larger sample (i.e. more studies investigating criterion validity) would be necessary for such a meta-analysis to provide a reliable result. Further investigation into the criterion validity of the IIEF-15 and IIEF-5 is necessary for a more nuanced interpretation.

More information is necessary regarding the measurement error of both the IIEF-15 and the IIEF-5. Currently, the only available evidence is based on one study of inadequate quality [86]. This evidence showed an insufficient value for the Erectile Function subscale,

but it is not possible to determine whether this is an artefact of the poor methodology of the study. Given the high frequency of use of both the IIEF-15 (particularly the Erectile Function subscale) and the IIEF-5 in clinical screening for ED, as well as outcome measures for clinical trials, knowledge on measurement error is important. to be able to know whether clinical change (i.e. clinical improvement or deterioration) is true change or whether it is an artefact of the measurement tool itself. Fortunately, one study of very good quality calculated the MCID using multiple methods on a very large sample [88]. This information can be used to interpret any measurement error that is calculated for the Erectile Function subscale. We recommend researchers performing a test-retest reliability design to calculate the Limits of Agreement or Smallest Detectable Change, to further inform the field. More studies investigating the MCID are also necessary to further interpret measurement error.

A limitation of this review is that we did not investigate content validity. Content validity needs to be established before other measurement properties can be regarded [38]. A future investigation of content validity is warranted. Another limitation of this review is the use of a precise rather than a sensitive search filter of measurement properties to identify studies to be included. The sensitivity of the precise filter was 93% in a random set of PubMed records, while the sensitivity of the sensitive search filter was 97% [61]. The use of the precise filter was a pragmatic choice over the available sensitive filter as the initial search encompassed 39 PROMs (including the IIEF-15 and IIEF-5), and the sensitive filter would provide too many hits for feasible screening. The possibility remains that the precise filter missed validation studies of the IIEF-15 and IIEF-5.

In 2002, the IIEF-15 was considered to "meet psychometric criteria for test reliability and validity" [60]. We offer a more cautious interpretation of the measurement properties of the IIEF-15. While we support the claim that the IIEF-15 meets psychometric criteria for criterion validity (in regard to the Erectile Function subscale) and responsiveness; we argue that structural validity, internal consistency, test-retest reliability, construct validity, and measurement error have not yet been demonstrated to meet psychometric criteria. Given the widespread of use of the IIEF-15 in both clinical practice and research, more thorough research is necessary regarding these measurement properties. A large-scale cross-cultural study design or an individual patient data meta-analysis, applying Confirmatory Factor Analysis, measurement invariance tests, internal consistency measures, and calculating the Limits of Agreement and/or Smallest Detectable Change, is recommended. It is possible that such research may suggest adjustments to be made to the IIEF-15 or its' scoring.

The results of this review highlight a couple of important points for the interpretation of the IIEF-15 and IIEF-5 in clinical practice and research. Firstly, some of the subscales may need to be combined and interpreting them as two separate constructs may not be valid. As the erectile function subscale is most often found in one factor with other subscales (based on both CFA and PCA), further research may find that other subscales should be combined with this subscale for a valid interpretation. Secondly, there is uncertainty what the optimal cut-off should be for the IIEF-15 and IIEF-5 to screen for ED, as multiple optimal cut-off scores were reported for both the IIEF-15 and IIEF-5. Further research is necessary to investigate optimal cut-off points. For current practice, it is important that researchers and clinicians maintain consistency, and as such the cut-off points of 25 for the IIEF-15 EF domain and 21 for the IIEF-5 should be maintained. We do suggest researchers and clinicians keep a close eye on further research of criterion validity, as another cut-of point may arise to be more accurate. Thirdly and lastly, the lack of information on measurement error is a problem for the interpretation of change scores of the IIEF-15 and IIEF-5. We advise to use the IIEF in tandem with another measure when determining ED development in patients, as this may lead to a more robust interpretation of change over time.

2.4 Conclusions

The IIEF-15 meets psychometric criteria for criterion validity (in regard to the Erectile Function subscale) and responsiveness; but structural validity, internal consistency, test-retest reliability, construct validity, and measurement error have not yet been demonstrated to meet psychometric criteria. In particular, further research into the structural validity of the IIEF-15 is of relevance. The IIEF-5 meets psychometric criteria for structural validity, test-retest reliability, construct validity, and criterion validity. Internal consistency, measurement error, and responsiveness require further research. The most pressing matter for future research is determining the unidimensionality of the IIEF-5.


3

Chapter 3

The measurement properties of the FSFI

Neijenhuijs, K. I., Hooghiemstra, N., Holtmaat, K., Aaronson, N. K., Groenvold, M., Holzner, B., … & Verdonck-de Leeuw, I. M. (2019). The Female Sexual Function Index (FSFI)—A Systematic Review of Measurement Properties. *The Journal of Sexual Medicine, 16(5),* 640-660. doi: 10.1016/j.jsxm.2019.03.001.

Abstract

Introduction: The Female Sexual Function Index (FSFI) is a patient reported outcome measure (PROM) measuring Female Sexual Dysfunction (FSD). The FSFI-19 was developed with six theoretical subscales in 2000. In 2010, a shortened version became available (FSFI-6). The current systematic review investigates the measurement properties of the FSFI-19 and FSFI-6.
Methods: A systematic search was performed of Embase, Medline, and Web of Science for studies that investigated measurement properties of the FSFI-19 or FSFI-6 up to April 2018. Data were extracted, and analysed according to COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) guidelines. Evidence was categorized into sufficient, insufficient, inconsistent, or indeterminate, and quality of evidence as very high, high, moderate, low.
Main Outcome Measures: The Main Outcome Measure is the evidence of a measurement property, and the quality of evidence based on the COSMIN guidelines.
Results: Eighty-three studies were included. Concerning the FSFI-19, the evidence for internal consistency was sufficient and of moderate quality. The evidence for reliability was sufficient but of low quality. The evidence for criterion validity was sufficient and of high quality. The evidence for structural validity was inconsistent of low quality. The evidence for construct validity was inconsistent of moderate quality. Concerning the FSFI-6, the evidence for criterion validity was rated as sufficient of moderate quality. The evidence for internal consistency was rated as indeterminate. The evidence for reliability was inconsistent of low quality. The evidence for construct validity was inconsistent of very low quality. No information was available on structural validity of the FSFI-6, and measurement error, responsiveness, and cross-cultural validity of both FSFI-6 and FSFI-19.
Clinical implications: Conflicting and lack of evidence for some of the measurement properties of the FSFI-19 and FSFI-6, indicates the importance of further research on the validity of these PROMs. We advise researchers whom use the FSFI-19 to perform confirmatory factor analyses and report the factor structure found in their sample. Regardless of these concerns, the FSFI-19 and FSFI-6 have strong criterion validity. Pragmatically, they are good screening tools for the current definition of FSD.
Strengths & Limitations: A strong point of the review is the use of pre-defined guidelines. A limitation is the use of a precise rather than a sensitive search filter.
Conclusions: The FSFI requires more research on structural validity (FSFI-19 and FSFI-6), reliability (FSFI-6), construct validity (FSFI-19), measurement error (FSFI-19 and FSFI-6), and responsiveness (FSFI-19 and FSFI-6). Further corroboration of measurement invariance (both across cultures and across subpopulations) in the factor structure of the FSFI-19 is necessary, as well as tests for the unidimensionality of the FSFI-6.

Sexual dysfunction refers to a problem that prevents people experiencing satisfaction from sexual activity. A first model of female sexual dysfunction (FSD) was composed in 1998 with four categories of disorders (in desire, arousal, orgasm, and pain) as described in the Diagnostic and Statistical Manual of Mental Disorders-IV (DSM-IV) and the International Statistical Classification of Diseases and Related Health Problems-10 (ICD-10) at that time [106].

In 2000, the Female Sexual Function Index (FSFI) was developed to measure female sexual (dys)function [54], based on the models described in the DSM-IV and ICD-10. The FSFI is a 19-item patient reported outcome measure (PROM), consisting of six separate domains of female sexual function, namely desire (items 1-2), arousal (3-6), lubrication (7-10), orgasm (11-13), satisfaction (14-16), and pain (17-19). Initial validation showed good internal consistency for all scales in a study sample drawn from the general population, as well as in subgroups of FSD patients and controls (Cronbach's Alpha = .82 - .97). Test-retest reliability was acceptable (r = .79 to .86). Known-groups comparison was tested between FSD patients and controls, with significant differences on all domains of the FSFI-19. Divergent validity (as measured with the Locke-Wallace Marital Adjustment Test [107]) was good (r = .04 to .43), except for the FSFI satisfaction scale (r = .40 to .72) [54]. In 2010, a 6-item version (FSFI-6) to measure FSD was developed. The six items were selected by inspecting the Receiver Operating Characteristic curves of each item of the FSFI-19 for distinguishing between women with and without FSD. The best-performing item for each of the six domains of the FSFI-19 was selected for use in the FSFI-6 [55]. The FSFI-6 showed acceptable internal consistency (Cronbach's Alpha = .789), acceptable test-retest reliability (Pearson correlation = .95), and good criterion validity with a cut-off of ≤ 19 (sensitivity = .96; specificity = .91).

With the release of the DSM-5 in 2013, the model for FSD has seen some changes. Of particular interest, one desire disorder (sexual aversion disorder) was removed, while the remaining desire disorder (hypoactive desire disorder) was merged with the arousal dysfunction disorder [108]. This new model suggests that desire and arousal may not be separate constructs in the context of FSD. Interestingly, the original validation study found a five-factor structure where desire and arousal were part of the same construct. This factor was split into two subscales due to clinical considerations [54].

The FSFI-19 and FSFI-6 are widely used in clinical practice as a screening tool for FSD, as well as in clinical trials as an outcome measure. As such, it is of importance to assess the measurement properties of the FSFI-19 and FSFI-6, to determine whether they are fit to use in clinical and scientific contexts. To our knowledge, the measurement properties of the FSFI-19 and FSFI-6 have not yet been systematically reviewed. As such, the aim of

this study was to investigate whether the initial good results regarding the measurement properties of the FSFI-19 and FSFI-6 were confirmed in later studies. Of particular interest is structural validity, and the question is whether the original six-factor structure is challenged in favour of a five-factor structure where desire and arousal are part of the same construct. The results of this systematic review are relevant for the use of the FSFI to monitor sexual dysfunction in females in clinical trials and practice.

In this review, we followed the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) methodology [38]. This methodology is based on a taxonomy and definitions of measurement properties for PROMs [37] including content validity (assessment of whether the FSFI represents all facets of FSD), structural validity (assessment of whether the FSFI subscales are singular constructs), internal consistency (assessment of whether FSFI items measuring the same construct are consistent in their results), cross-cultural validity (assessment of whether there are structural differences in validity of the FSFI between populations), reliability (assessment of whether the FSFI reproduces similar scores when FSD has been stable), measurement error (assessment of systematic and random error between the FSFI score and the true score of a patient), criterion validity (assessment of how well the FSFI score is an adequate reflection of FSD diagnosis), hypotheses testing for construct validity (assessment of whether the FSFI measures the construct of FSD), and responsiveness (assessment of how well the FSFI measures FSD change over time).

3.1 Methods

3.1.1 Literature search

The literature search was part of a larger systematic review (Prospero ID 42017057237), which investigated the measurement properties of 39 different PROMs (including the FSFI) measuring quality of life of cancer survivors included in an eHealth application called "Oncokompas" [8–11]. The databases Embase, Medline, and Web of Science were searched using the search terms of the PROM's name and acronyms, combined with a precise search filter for measurement properties [61]. The search was performed in January 2017. A search update was performed on April 13th 2018, to search for recent studies. This search update also used broader search terms across all years (not only from 2017 to 2018) as not all acronyms of the FSFI were correctly specified in the original search. Appendix A contains the full search terms. Due to the limitation of the sensitivity of the precise filter (93% sensitive) [61], a manual search using rudimentary search filters was performed in Google Scholar and Pubmed to check for any prominent records missed in the search update.

3.1.2 Inclusion and exclusion criteria

Studies were included when they reported original data on at least one of the following measurement properties of the FSFI: structural validity (whether the hypothesized measurement model is confirmed), internal consistency (the degree of interrelatedness among the items of the measure), reliability (the proportion of total variance between multiple measurements which is due to "true" differences between measurements), measurement error (a measure of systematic and random error in change scores), criterion validity (whether the measure is an adequate reflection of a gold standard; in the case of the FSFI this is most often a diagnosis of FSD), cross-cultural validity (whether the test can be interpreted similarly in different cultures), responsiveness (whether the measure is capable of measuring change over time in the construct to be measured), and hypothesis testing for construct validity (whether the test measures the construct it proposes to measure) which consists of known-groups comparison (a comparison between groups known to have differences on the construct), convergent validity (correlations with other measures that should be related), and divergent validity (correlations with other measures that should be unrelated). While of importance for establishing validity, content validity was not investigated as it was beyond the scope of the current review. Validation studies on other PROMs that also reported original data on the FSFI were included as well.

Studies that were only available as abstracts or conference proceedings were excluded, as well as non-English publications. Titles and abstracts, and the selected full-texts were screened by two independent reviewers (KN / MV / KH / NH). Disagreements were discussed until consensus was reached.

3.1.3 Data extraction

Data on each of the measurement properties defined by the COSMIN taxonomy [37] was extracted by two independent researchers (KN / AvdH / HM / EV / NH). Relevant data included the type of measurement property, its' results, and information on missing values. Information on the type of research (psychometric or not), specified research aim, sample size, population information, and which version of the FSFI was used, was also extracted. Disagreements were discussed until consensus was reached.

3.1.4 Data analysis

Data analysis consisted of three consecutive steps. First, the quality of the included studies was rated using the 4-point scoring system of the COSMIN checklist [64]. Methodological aspects regarding design requirements and preferred statistical methods, specific to each measurement property under consideration, were rated as either "inadequate", "doubtful", "adequate", or "very good". The methodological quality was summarized per measurement property per study, as the lowest score received on any of the methodological aspects. The

complete criteria for study quality per measurement property are documented elsewhere [64]. Appendix E contains the final study quality ratings.

Second, each measurement property in each individual study was rated as sufficient, insufficient or indeterminate, according to criteria for good measurement properties included in the COSMIN guidelines for systematic reviews of PROMs. The complete criteria for rating these measurement properties are documented elsewhere [38]. These ratings were qualitatively summarized to determine the overall rating of the measurement property for the FSFI. If all studies indicated a "sufficient", "insufficient", or "indeterminate" rating for a specific measurement property, the overall rating of this measurement property was rated accordingly. If there were inconsistencies between studies, explanations were explored (e.g. differences in methodological quality, differences in population, etc.). If explanations were found, they were discussed until consensus was reached regarding the overall rating of the measurement property. If no explanations were found, the overall rating would be inconsistent.

Third, the overall rating of evidence per measurement property was supplemented by a level of quality of evidence, using a modified Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach from the COSMIN methodology [38]. This approach takes into account (i) study quality, (ii) directness of evidence, (iii) inconsistency of results, and (iv) precision of evidence (number of studies and sample size). The overall quality of evidence was rated as high, moderate, low, or very low. Measurement properties that were rated as indeterminate in the previous step, did not receive a rating in this third step as there was no evidence to rate.

All ratings (methodological quality, measurement property rating, and GRADE rating) were made by two independent researchers (KN / KH / NH). Discrepancies in ratings were discussed until consensus was reached.

3.2 Results

3.2.1 Search results

The initial search identified 1401 non-duplicate abstracts of which 174 were relevant to the FSFI (*Figure 3.1*). A total of 155 abstracts and 11 full-texts were excluded from the initial search, as they did not provide unique information on a measurement property. The search update up to April 2018 identified 1415 more non-duplicate abstracts. A total of 1229 abstracts and 110 full-texts were excluded from the search update, as they did not provide unique information on a measurement property of the FSFI. Two full-texts were excluded during data-extraction.

In total we included eighty-three studies: seventy-five on the FSFI-19 [54, 73, 81, 85, 104, 109–178], five on the FSFI-6 [55,179–182], and three on adaptations of the FSFI-19: a version specific for breast cancer survivors [183], a version for life-long sexual dysfunction [184], and a version with an added item concerning vaginismus [185]. An overview of study characteristics is provided in Table 3.1.

The combined body of the seventy-five studies on the FSFI-19, and the five studies on the FSFI-6 reported on all measurement properties, except measurement error, responsiveness, and cross-cultural validity. The three studies on adaptations of the FSFI-19 reported on structural validity, internal consistency, test-retest reliability, and construct validity of original subscales of the FSFI-19.

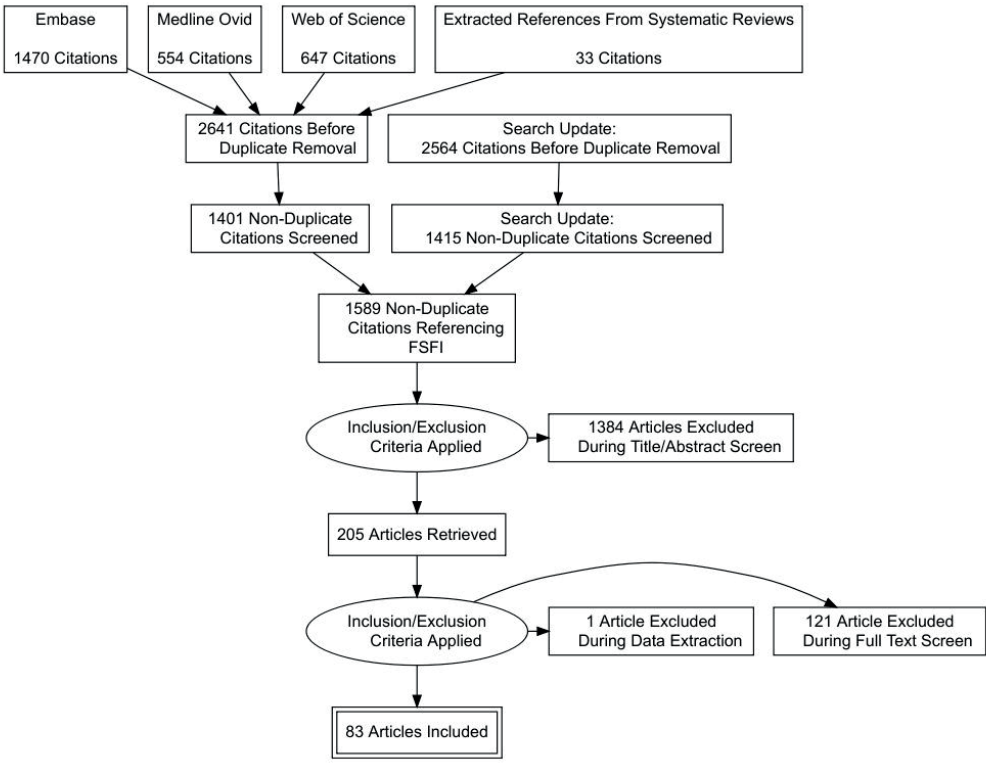

Figure 3.1: PRISMA diagram.

Table 3.1. Characteristics of included studies.

Reference	Population	Sample size	Main aim of study
FSFI-19			
Achimas-Cadariu et al. (2013) [109]	Female patients of reproductive age treated for premalignant and malignant pathology of the uterine cervix	102/204	To investigate the associations among quality of life determinants on a sample of female patients of reproductive age (102 patients and 102 healthy controls), surgically treated (conisation) for pre-invasive and invasive pathology (cervix dysplasia and micro invasive carcinoma)
Ahmed et al. (2017) [110]	Premonopausal women	241	To assess sexually related personal distress among premenopausal women with sexual dysfunction via a validated Arabic version of the original FSDS
Anis et al. (2011) [111]	Egyptian women	855	To validate the Arabic version of the Female Sexual Function Index (ArFSFI)
Aydin et al. (2016) [112]	Turkish women visiting the urogynecology clinic	248	To develop a Turkish version of the FSDS-R, to evaluate its psychometric reliability and validity, and to estimate the optimal cutoff score that corresponds best to the clinical diagnosis of sexual dysfunction
Azimi Nekoo et al. (2014) [113]	Married and potentially sexually active Iranian women	1966	to determine the psychometric properties of the Iranian version of the FSDS-R in a population sample of Iranian women
Bartula et al. (2015) [114]	Breast cancer patients	399	Assess extent to which FSFI is applicable for breast cancer patients
Baser et al. (2012) [115]	Female cancer survivors	181	Systematic evaluation of the factor structure, reliability, and construct validity of the FSFI for measuring the sexual functioning of female cancer survivors
Bloemendaal et al. (2015) [116]	Dutch women	323	Validate the Dutch version of the Sexual Excitation/Sexual Inhibition Inventory for Women
Borello-France et al. (2008) [117]	Female patients with a relapsing form of MS	48	To assess the test-retest reliability of the Urge-Urinary Distress Inventory (U-UDI) and the Female Sexual Function Index (FSFI) in women with MS
Bornefeld-Ettmann et al. (2018) [118]	German-speaking women	465	The German translation of the Sexual Self-Esteem Inventory-Short Form (SSEI-SF) by Zeanah and Schwarz (1996) was validated via an online survey with 557 women and then investigated in a clinical sample of women suffering from PTSD following sexual and physical abuse compared with healthy controls
Burri et al. (2018) [119]	Sexually active Swiss women	309	To evaluate the validity and utility of the German version of the SCS-W by assessing content, convergent, and discriminant validity
Carpenter et al. (2015) [120]	Midlife postmenopausal women	93	To evaluate whether a single item from the FSDS-R could be identified to use to screen midlife women for bothersome diminution in sexual function

Reference	Population	Sample size	Main aim of study
Carpenter et al. (2016) [121]	Peri- and postmenopausal women reporting hot flashes	898	To evaluate whether a subset of items on the 19- item English-language FSFI would perform as well as the full length FSFI in peri- and postmenopausal women
Chang et al. (2009) [122]	Pregnant women receiving prenatal examinations	108	To translate the Female Sexual Function Index (FSFI) from English to traditional Chinese, and to evaluate the reliability and validity of this new version for pregnant women
Clayton et al. (2006) [123]	Female patients with diagnosis of HSDD	90	To assess the reliability and validity of the SIDI-F as a measure of HSDD severity
Clayton et al. (2010) [124]	American (N=220) and European (N=253) women going through FSD diagnosis	473	To estimate the reliability and validity of the SIDI-F as a measure of HSDD severity
Constantine et al. (2017) [125]	American and British women	589	To create a valid and responsive summary score for the PISQ-IR
DeRogatis et al. (2010) [126]	Postmenopausal emale patients aged 40-65 with spontaneous amenorrhea or bilateral oophorectomy with or without hysterectomy at least 6 months prior to study	629	To validate the WSID-SF and DSLA in postmenopausal women
Eaton et al. (2017) [127]	Female cancer survivors	175	To develop and validate brief clinical measurements to facilitate the identification of vulvovaginal symptoms in patients with and survivors of cancer
Fakhri et al. (2012) [128]	Iranian gynecological outpatients	448	To translate, validate, and enhance cross-cultural comparability of an Iranian version (IV) of the Female Sexual Function Index
Farkas et al. (2016) [129]	Female patients diagnosed with Pelvic Organ Prolapse, Urinary Incontinence, or Fecal Incontinence	178	To translate the Prolapse/Urinary Incontinence Sexual Questionnaire, IUGA-Revised (PISQ-IR), into Hungarian and to validate the translated PISQ-IR
Ferguson et al. (2012) [130]	Women visiting gynecologic oncology outpatient clinic	268	To confirm the factor structure of the Sexual Adjustment and Body Image Scale using a confirmatory factor analysis
Filocamo et al. (2014) [131]	Italian women visiting urological and gynecological clinics	409	To perform a linguistic validation of the Italian version of the FSFI
Flynn et al. (2013) [73]	American female cancer patients	430	Validation of PROMIS sexual function and satisfaction scales
Forbes et al. (2014) [104]	Sexually active Australian women	336	To examine the measurement capabilities of the IIEF and FSFI based on data collected from an online study in 2010
Ghassamia et al. (2013) [132]	Iranian women	562	To examine the psychometric properties of a Persian language version of the Female Sexual Function Index (P-FSFI) amongst a sample of healthy Iranian women
Heng et al. (2013) [134]	Malaysian women visiting infertility clinic	150	To determine the construct of the phases of the female sexual response cycle (SRC) among women attending an infertility clinic in a Malaysian tertiary center

Reference	Population	Sample size	Main aim of study
Herbenick et al. (2010) [135]	American women attending sex toy parties	1937	To establish a reliable and valid measure of female genital self-image, the Female Genital Self-Image Scale (FGSIS), and to assess the relationship between scores on the FGSIS and women's sexual function
Herbenick et al. (2011) [136]	American women	2056	To assess the reliability and validity of the FGSIS, its model of fit, and its association with women's scores on the Female Sexual Function Index (FSFI) in a nationally representative probability sample of women in the United States ages 18 to 60
Hevesi et al. (2017) [137]	202 university students, 177 patients with endometriosis, and 129 patients with polycystic ovary syndrome; from Hungary	508	To investigate whether female sexual function is best understood as a multidimensional construct or, alternatively, whether a common underlying factor explains most of the variance in FSFI scores
Ismail et al. (2014) [138]	178 female patients with diabetes, and 175 women without diabetes from Malaysia	353	To compare the components of sexual responses between Malaysian women with Type 2 diabetes mellitus and those without the disease
Jing et al. (2018) [139]	Breast cancer survivors	246	To develop a Quality of Sexual Life Questionnaire in Breast Cancer Surivivors and determine its validity and reliability
Kalmbach et al. (2015) [140]	Female undergraduate students	409	To assess factor structures of the Female Sexual Function Index (FSFI), Male Sexual Function Index (MSFI) (adapted for this investigation), and Profile of Female Sexual Function (PFSF) in young, healthy men and women
Likes et al. (2006) [141]	43 female patients with vulvar excisions for vulvar intraepithelial neoplasia; 43 age-matched controls	86	To extend the validation of the Female Sexual Function Index to include women with vulvar excisions for vulvar intraepithelial neoplasia
Liu et al. (2014) [142]	Chinese female patients with interstitial cystitis and bladder pain syndrome	90	To examine whether adding a sexual dysfunction domain to urinary, psychosocial, organ specific, infection, neurologic or systemic, and tenderness (UPOINT) system improves the association with interstitial cystitis and bladder pain syndrome (IC-BPS) symptom severity due to a high prevalence of sexual dysfunction in women
Liu et al. (2016) [143]	Female inpatients with cervical cancer	215	To examine the psychometric properties and performance of a Chinese version of the Female Sexual Function Index (FSFI) among a sample of Chinese women with cervical cancer
Ma et al. (2014) [144]	Chinese women	500	To establish clinical cutoff scores for the CVFSFI and to evaluate the prevalence of FSD in urban Chinese women
Meston et al. (2003) [145]	71 female patients with female orgasmic disorder, 44 female patients with hypoactive sexual desire disorder, and 71 healthy women	186	To extend the validation of the FSFI to include women with a primary clinical diagnosis of female orgasmic disorder or hypoactive sexual desire

Reference	Population	Sample size	Main aim of study
Meston et al. (2005) [146]	American women	172	To develop a comprehensive, multifaceted, valid, and reliable self-report measure of women's sexual satisfaction and distress
Mestre et al. (2017) [147]	118 not sexually active women, and 150 sexually active women	268	To transculturally adapt the Pelvic Organ Prolapse/Incontinence Sexual Questionnaire IUGA-Revised (PISQ-IR) into Spanish
Mohammadi et al. (2014) [148]	Iranian married women with MS	226	To translate and validate the MSISQ-19 in women with MS in Iran
Mohammed et al. (2014) [149]	Egyptian married women	244	To translate the original English version of the Female Genital Self-Image Scale (FGSIS) into Arabic and validate the Arabic version
Nimbi et al. (2018) [81]	Italian women	626	To test the psychometric characteristics of the Italian version of the SMQ focusing on the Automatic Thoughts subscale
Nowosielski et al. (2013) [150]	85 Polish female patients with FSD, 104 Polish women without FSD	189	To develop a Polish version of the FSFI
Opperman et al. (2013) [151]	Canadian women	85	To evaluate and compare four models of the Female Sexual Functioning Index: (a) single-factor model, (b) six-factor model, (c) second-order factor model, and (4) five-factor model combining the desire and arousal subscales
Pakpour et al. (2013) [152]	Iranian female population sample (N=2675), Iranian female patients with FSD (N=295), Iranian female patients with type 2 diabetes (N=449)	3419	The purpose of this study was the translation and validation of an Iranian version of the Sexual Quality of Life questionnaire-Female (SQOL-F) in Iranian women
Pakpour et al. (2014) [153]	Iranian female students	1877	To investigate the psychometric properties of a translated and culturally adapted Iranian version of the Female Genital Self-Image Scale (FGSIS-I) in a sample of college women
Pascoal et al. (2017) [85]	Heterosexual sexually active women involved in a dyadic relationship	278	To describe the development and validation of the Beliefs About Sexual Functioning Scale
Rehman et al. (2015) [154]	Bilingually educated women in a stable sexual relationship	116	To translate, cross-culturally adapt, and perform a psychometric validation of an Urdu translation of the Female Sexual Function Index
Rellini et al. (2006) [155]	Female patients with female sexual arousal disorder	24	To provide empiric evidence on the sensitivity of different types of measures for detecting treatment-induced changes in female sexual dysfunction diagnosis
Rillon-Tabil et al. (2013) [156]	Ambulatory women	85	To translate and validate the Female Sexual Function Index Filipino version
Rogers et al. (2013) [157]	American and British female patients with pelvic floor disorders	589	To create a valid, reliable, and responsive sexual function measure in women with pelvic floor disorders (PFDs) for both sexually active (SA) and inactive (NSA) women
Rosen et al. (2000) [54]	Healthy women	259	To develop and psychometrically validate a self-administered Female Sexual Well-Being Scale™ for assessing sexual well-being in sexually functional women

Reference	Population	Sample size	Main aim of study
Rosen et al. (2009) [158]	American women reporting normal sexual function	329	Identifying a giagnostic cut-point for differentiating women with and without HSDD
Ryding et al. (2015) [159]	50 Swedish female patients with hypoactive sexual desire disorder, and 58 age-matched healthy Swedish women	108	To investigate the psychometric properties of the Swedish version of the FSFI
Selcuk et al. (2016) [160]	71 Turkish female patients with perlvic problems, and 38 Turkish healthy women	109	To validate the Turkish versions of the SHOW-Q for Turkish-speaking women
Sidi et al. (2007) [161]	Married Malaysian women	230	To validate the Malay version of the Female Sexual Function Index
Sills et al. (2005) [162]	Premenopausal female patients diagnosed with HSDD	448	To use the outcome of item response analyses of blinded data from two randomized, placebo-controlled trials, to assist in the revision of the scale
Stephenson et al. (2016) [163]	Adult American women in a monogamous heterosexual relationship reporting sexual difficulties	97	To assess the correlations between FSFI scores and information regarding specific rates of functional impairment gained via clinical interview; and (b) to assess the specificity of FSFI subscale scores in reflecting corresponding aspects of sexual function
Sun et al. (2011) [164]	85 Chinese women seeking regular health check-up, 145 Chinese women who accompanied patients, and 98 Chinese female patients with medical illness not affecting sexual function	328	To develop and validate the Chinese version of the Female Sexual Function Index (CVFSFI) to assess FSD in China
Takahashi et al. (2011) [165]	Healthy Japanese women in partnered relationships	126	To develop a Japanese translation of the Female Sexual Function Index (FSFI-J), 3-month version, and to measure its psychometric reliability and validity
Ter Kuile et al. (2006) [166]	234 Dutch female patients with FSD, and 108 Dutch women without FSD	342	To investigate the psychometric properties of the Female Sexual Function Index and the Female Sexual Distress Scale within a Dutch population
Trudel et al. (2012) [167]	Canadian women >65 years old in a relationship	143	To validate the FSFI in an older (65 years and over), non-clinical population of francophone women living with their spouses in Quebec
Trutnovsky et al. (2016) [168]	German female patients visiting urogynecological clinics for pelvic floor dysfunction	197	To translate the Pelvic Organ Prolapse/ Incontinence Sexual Questionnaire–International Urogynecology Association Revised (PISQ-IR) into German and to clinically validate it in a German-speaking population
Vallejo-Medina et al. (2018) [169]	Colombian adult women	925	To adapt and validate the FSFI to Spanish language in a Colombian sample
Velten et al. (2016) [170]	German adult women	2206	To assess the psychometric properties of a German version of the SESII-W
Verit et al. (2007) [171]	100 female patients with CPP and 100 age-matched women without CPP	200	To investigate the validity and reliability of Female Sexual Function Index (FSFI) in women with Chronic Pelvic Pain

Reference	Population	Sample size	Main aim of study
Wang et al. (2015) [172]	Chinese women visiting a urogynecological clinic	106	To translate and validate the Mandarin Chinese version of PISQ-IR for global use
Wiegel et al. (2005) [173]	307 female patients with FSD diagnoses, and 261 healthy women	568	To cross- validate the FSFI in several samples of women with mixed sexual dysfunctions (N = 568) and to develop diagnostic cut-off scores for potential classification of women's sexual dysfunction
Witting et al. (2008) [174]	Finnish female adult twins	2081	To validate the FSFI in Finnish
Wolpe et al. (2017) [175]	Brazilian female physical therapy students	246	To assess the psychometric properties of the FSFI applied to the VAS
Wylomanski et al. (2014) [176]	French women attending gynecology consultation	512	To validate a French version of the Female Sexual Function Index (FSFI) in a sample of French women
Zachariou et al. (2017) [177]	18 Greek female patients with FSD, and 99 Greek women without FSD	117	To linguistically validate the Greek version of Female Sexual Function Index
Zohre et al. (2014) [178]	100 Iranian healthty women, 200 Iranian female patients suffering from Urinary Incontinence with or without Pelvic Organ Prolapse	200	To translate the Pelvic Organ Prolapse/Urinary Incontinence Sexual Questionnaire (PISQ-12) and provide evidence for psychometric properties
FSFI-19 Sexual Desire subscale only			
Gerstenberger et al. (2010) [133]	American and Canadian women	618 & 892	To define and validate a specific cut point on the SD domain for differentiating women with and without hypoactive sexual desire disorder
FSFI-19; FSFI-LL			
Burri et al. (2010) [184]	British female twins	FSFI-19: 1056; FSFI-LL: 744	To develop a modified version of the widely used FSFI which allows assessment of women's lifelong sexual function—the FSFI-LL—and to evaluate the psychometric properties and aptness of this new version
FSFI-20 (added item for vaginismus)			
Carvalho et al. (2012) [185]	Portuguese women	1425	To test, using structural equation modeling, five conceptual, alternative models of female sexual function, using a sample of women with sexual difficulties and a sample of women without sexual problems
FSFI-BC (34 items)			
Bartula et al. (2015b) [183]	Australian breast cancer survivors	596	To determine the reliability, validity, and acceptability of a breast cancer-specific adaptation of the Female Sexual Function Index, the FSFI-BC
FSFI-6			
Chedraui et al. (2012) [179]	Ecuadorian women	904	To assess sexual function and related factors in mid-aged Ecuadorian women

Reference	Population	Sample size	Main aim of study
Isidori et al. (2010) [55]	Women attending a screening visit for sexual and reproductive dysfunctioning	160	Development of short-form version of FSFI
Lee et al. (2014) [180]	Korean female patients who visited outpatient center for uterine cancer	220	To evaluate the validity and reliability of the Korean version of the Female Sexual Function Index-6 (FSFI-6K)
Mitchell et al. (2012) [181]	1262 population sample, and 100 patients with sexual problems. Count of women unspecified.	1362	Development of a new measure of sexual function for the third British National Survey of Sexual Attitudes and Lifestyles
Pérez-López et al. (2012) [182]	Female patients attending gynecological and obstetrical healthcare facilities	179	To assess sexual function and related factors in mid-aged Spanish women

FSFI: Female Sexual Function Index; HSDD: Hypoactive Sexual Desire Disorder; FSD: Female Sexual Disorder; CPP: Chronic Pelvic Pain

3.2.2 Structural validity

Twenty-nine studies reported on structural validity of the FSFI-19 [54, 104, 111, 114, 115, 119–122, 128, 132, 134, 137, 138, 140, 143, 150, 151, 154, 156, 164–166, 169, 173–176, 183, 184], of which eight reported multiple analyses [54, 128, 143, 166, 169, 174, 176, 184] (*Table 3.2*). Methodological quality of these studies was rated as "very good" [54, 114, 137, 140], "adequate" [104, 111, 115, 119, 120, 122, 128, 132, 134, 138, 143, 150, 164–166, 169, 173–176, 183, 184], or "inadequate" [54, 151, 156]. The "inadequate" ratings were due to sample sizes that were too small ("other flaws" in COSMIN methodological quality).

Three studies of "very good" quality [114, 137, 184], one studies of "adequate" quality [169], and one study of "inadequate" quality [151] confirmed the hypothesized six-factor structure, and were thus rated as sufficient. Eight studies of "adequate" quality [120, 128, 140, 143, 165, 169, 174, 176] showed a poor fit for the six-factor structure. Two of these studies tested and showed support for a five-factor structure [128, 143].

Nineteen studies of "adequate" quality [104, 111, 115, 119, 122, 128, 132, 134, 138, 143, 150, 154, 164, 166, 174–176, 183, 184], and two studies of "inadequate" quality [54, 156] performed analyses (mostly Principal Component Analysis) without reporting fit statistics, therefore results were rated as indeterminate. Noteworthy is that five studies showed support for a six-factor structure [111, 156, 166, 173, 176], while eleven studies showed support for a five-factor structure with a merging of the desire and arousal subscales [54, 104, 115, 119, 128, 143, 150, 154, 173, 183, 184]. One study showed support for a five-factor structure with a different merging of subscales [132], and seven studies showed support for less than five factors [122, 134, 138, 164, 173–175]. One study used Item Response Theory analysis and was rated indeterminate as no fit measures were reported [121].

None of the studies that investigated the FSFI-6 [55,179–182] reported on structural validity.

Table 3.2. Structural validity of the FSFI.

Reference	Methodology	Outcome	Rating	Quality
FSFI-19				
Anis et al. (2011) [111]	PCA	Six-component structure	Indeterminate	Adequate
Bartula et al. (2015) [114]	CFA	Six factors with item 14 removed: desire, arousal, lubrication, orgasm, satisfaction, pain	Sufficient	Very good
Bartula et al. (2015b) [183]	PCA	Five components: desire/arousal, lubrication, orgasm, satisfaction, pain	Indeterminate	Adequate
Baser et al. (2012) [115]	EFA	Five factors: desire/arousal, lubrication, orgasm, satisfaction, pain	Indeterminate	Adequate
Burri et al. (2010) [184]	PCA	FSFI-19: Unrotated PCA identified five components. Although the sixth component had a considerably low eigenvalue, subsequent equamax rotation yielded the most consistent pattern of factor loadings using a six-component structure. FSFI-LL: Unrotated PCA identified three components. Although the fourth and fifth components had considerably low eigenvalues, subsequent equamax rotation yielded the mosts consistent pattern of factor loadings using a five-component structure.	Indeterminate	Adequate
Burri et al. (2010) [184]	CFA	FSFI-19: The six-factor solution was acceptable after allowing correlations between subscales and between a number of items. FSFI-LL: The five-factor solution was acceptable after allowing correlations between subscales, and between a number of items.	Sufficient	Adequate
Burri et al. (2018) [119]	PCA	Unrotated PCA identified 5 components with eigenvalues higher than 1. Although the 6th component had an eigenvalue <1, subsequent varimax rotation yielded the most consistent pattern of factor loadings using a six-component structure.	Indeterminate	Adequate
Carpenter et al. (2016) [121]	IRT	After pruning based on violations of local independence, on discrimination, and difficulty parameters; 9 items remained of the 19 items of the FSFI-19.	Insufficient	Adequate
Chang et al. (2009) [122]	PCA	Three components were extracted and identified with eigenvalues greater than 1.03. These three components accounted for a total of 87.10% of the variance. Component 1, with an initial eigenvalue of 13.74, accounted for 72.32% of the explained variance. The three components were interpreted as "coitus" (15 items), "satisfaction" (2 items), and "desire" (2 items).	Indeterminate	Adequate

Reference	Methodology	Outcome	Rating	Quality
Fakhri et al. (2012) [128]	PCA	The PCA yielded a best fitting, five-component solution. All five components had eigenvalues of greater than one and accounted for 70% of the total variance.	Indeterminate	Adequate
Fakhri et al. (2012)[128]	CFA	Six-factor structure showed inadequate fit (chi2 = 826.60; df = 136; GFI = 0.72; CFI = 0.81; NNFI = 0.63; SRMR, 0.18; PNFI = 0.63; RMSEA = 0.15). Five-factor structure showed acceptable fit (chi2 = 304.07; df = 142; GFI = 0.89; CFI = 0.95; NNFI = 0.94; SRMR, 0.08; PNFI = 0.71; RMSEA = 0.07).	Insufficient	Adequate
Forbes et al. (2014) [104]	PCA	Five components with eigenvalues >1. The five components were clearly defined as desire and subjective arousal, lubrication, orgasm, satisfaction, and pain.	Indeterminate	Adequate
Ghassamia et al. (2013) [132]	PCA	Five components were extracted with eigenvalues >1. The examination of the scree plot suggested that four or five dimensions underlie the FSFI. The components were interpreted as "Sexual Response" (11 items), "Sexual-related Pain" (3 items), "Sexual Desire" (2 items), "Sexual Satisfaction" (3 items).	Indeterminate	Adequate
Heng et al. (2013) [134]	PCA	Three components were extracted with eigenvalues >1. The first component comprised sexual arousal, lubrication and pain. The second component comprised orgasm and sexual satisfaction. Sexual desire alone made the third component.	Indeterminate	Adequate
Hevesi et al. (2017) [137]	CFA	Six-factor model had an acceptable fit (Satorra-Bentler chi-square = 490.924, df = 137, p < .001; chi-square/df = 3.583; CFI = 0.960; TLI = 0.950; RMSEA = 0.071; range = 0.065-0.078). However, most intercorrelations among the factors were very high. A bi-factor model where each item was associated with a general factor and with its domain-specific factor showed an improvement from the original model (SatorraBentler chi-square = 272.630, df = 123, P < .001; chi-square/df = 2.217; CFI = 0.983; TLI = 0.976; RMSEA = 0.049; range = 0.041-0.057). It was found that in the total sample most observed variance was attributable to the general sexual function factor; while in the sexually active subsample most observed variance was attributable to the specific factors.	Sufficient	Very good
Ismail et al. (2014) [138]	PCA	Among the women without type 2 diabetes, three components were extracted with eigenvalues >1: Sexual desire/arousal, satisfaction, and pain. With the items in lubrication and orgasm domains loading on both satisfaction and pain. Among the women with Type 2 diabetes, three components were extracted with eigenvalues >1: A component comprising of lubrication, orgasm, and pain; satisfaction, and desire/arousal.	Indeterminate	Adequate

Reference	Methodology	Outcome	Rating	Quality
Kalmbach et al. (2015) [140]	CFA	Bad model fit of six-factor model (chi2 (137) = 683.28, p<.001, CFI = .91, TLI = .88, RMSEA = .07). Adding latent variables describing whether an item was positively or negatively worded increased the fit (chi2 (118) = 303.01, p<.001, CFI = .97, TLI=.95, RMSEA=.04).	Insufficient	Very good
Liu et al. (2016) [143]	PCA	Five components were extracted with an eigenvalue >1, accounting for 77.57% of the total variance. The first component consisted of a mixture of desire/arousal, and the rest were lubrication, orgasm, satisfaction, and pain.	Indeterminate	Adequate
Liu et al. (2016) [143]	CFA	A six-factor model showed a bad fit (CMIN/DF = 3.12, GFI = .83, CFI = .91, RMSEA = .100). A five-factor model showed an acceptable fit (CMIN/DF = 3.08, CFI = .91, GFI = .83, RMSEA = .099). The five factors included desire/arousal, lubrication, orgasm, satisfaction, and pain.	Insufficient	Adequate
Nowosielski et al. (2013) [150]	PCA	Five components were extracted, accounting for 83.62% of the total variance. The components reflected desire/arousal, lubrication, orgasm, satisfaction, and pain.	Indeterminate	Adequate
Opperman et al. (2013) [151]	CFA	A first-order, one-factor model showed a bad fit: chi2 (152, N = 85) = 664.45, p < .001 (chi2/df = 4.4; GFI = .55, TLI = .34, CFI = .41, and RMSEA = .20). A first-order, six-factor model with correlations among factors showed a good fit: chi2(137, N = 85) = 178.96, p = .009 (chi2/df = 1.3; GFI = .83, TLI = .94, CFI = .95, and RMSEA = .06). A second-order, six-factor model with one second-order factor showed a decrement in fit compared to the first-order, six-factor model: chi2(146, N = 85) = 199.72, p = .002 (chi2/df= 1.4; GFI = .80, TLI= .93, CFI= .94, and RMSEA = .066). A first-order, five-factor model showed a slight decrement in fit compared to the first-order and second-order, six-factor models: chi2(137, N = 85) = 215.89, p < .001 (chi2/df = 1.5; GFI = .79, TLI = .90, CFI = .92, and RMSEA = .079). Delta chi-square tests of differences indicated a better fit (Delta chi2(9, N = 85) = 21.62, p < .05) of the first-order, six-factor model (chi2(137, N = 85) = 178.96) versus the second-order, six-factor model (chi2(146, N = 85) = 199.72), as well as a significant better fit (Delta chi2(4, N = 85) = 37.79, p < .01) versus the first-order, five-factor model (chi2(142, N = 85) = 215.89).	Sufficient	Inadequate
Rehman et al. (2015) [154]	PCA	Five component structure: Desire, arousal, lubrication, orgasm, satisfaction and pain with eigenvalues 7.556, 3.457, 2.939, 2.926 and .633 respectively. These five components accounted for 92.164% of the explained variance.	Indeterminate	Adequate
Rillon-Tabil et al. (2013) [156]	PCA	Six-component structure	Indeterminate	Inadequate

Reference	Methodology	Outcome	Rating	Quality
Rosen et al. (2000) [54]	CFA	Five factors: desire/arousal, lubrication, orgasm, satisfaction, and pain; but desire/arousal was split into two factors due to theoretical considerations	Indeterminate	Very good
Rosen et al. (2000) [54]	PCA	Five components: desire/arousal, lubrication, orgasm, satisfaction, and pain	Indeterminate	Inadequate
Sun et al. (2011) [164]	PCA	Four components were extracted with eigenvalue >1. The first component was a mixture of arousal/orgasm/satisfaction and the remaining three components were lubrication, pain, and desire. These four components accounted for a total of 75.01% of the explained variance	Indeterminate	Adequate
Takahashi et al. (2011) [165]	EFA	Five-factor structure found: desire/arousal, lubrication, orgasm, satisfaction, and pain	Insufficient	Adequate
Ter Kuile et al. (2006) [166]	SCA	Six-component structure explained 88.6% variance.	Indeterminate	Adequate
Ter Kuile et al. (2006) [166]	PCA	Six-component structure explained 81.6% variance.	Indeterminate	Adequate
Vallejo-Medina et al. (2018) [169]	EFA	Five-factor structure found with a clear Arousal–Desire fusion.	Insufficient	Adequate
Vallejo-Medina et al. (2018) [169]	CFA	A six-factor uncorrelated factor model showed a bad fit (S-B chi2(df=146) = 550.02, p <.001, CFI = .976, RMSEA =.076, AIC = 258.02). A six-factor correlated factor model showed a good fit (S-B chi2(df=131) = 209.31, p <.001, CFI = .995, RMSEA =.036, AIC = -52.68). A second-order, six-factor model showed a good fit (S-B chi2(df=145) = 353.60, p <.001, CFI = .988, RMSEA =.055, AIC = 63.60). A five-factor correlated model showed a good fit (S-B chi2(df=137) = 338.86, p <.001, CFI = .988, RMSEA =.056, AIC = 64.86).	Sufficient	Adequate
Wiegel et al. (2005) [173]	PCA	A PCA on a sample of sexually functional and dysfunctional women (N=272) showed a five-component structure: desire/arousal, lubrication, orgasm, pain, and satisfaction. A PCA in women with sexual dysfunction, resulted in four components with eigenvalues >1 and one component with eigenvalue of .98. The four components were interpreted as: pain, orgasm, lubrication, desire/arousal/satisfaction. When taking the fifth component into account they were interpreted as: desire/arousal, lubrication, orgasm, pain, and satisfaction. A PCA in women without sexual dysfunction, resulted in five components with eigenvalues >1, which were interpreted as: desire/arousal, orgasm/arousal, lubrication, satisfaction, and pain. A PCA of the combined group (dysfunctional and nondysfunctional; N=527) resulted in five components with eigenvalues >1, which were interpreted as: desire/arousal, orgasm/arousal, lubrication, satisfaction, and pain.	Indeterminate	Adequate

Reference	Methodology	Outcome	Rating	Quality
Witting et al. (2008) [174]	EFA	Four factors had an eigenvalue > 1. The fifth factor had an eigenvalue of 0.84 for Twin 1 group and 0.85 for Twin 2 group. The sixth factor had an eigenvalue of 0.57 and 0.62, respectively. After exploring four, five, and six factor solutions, it was decided to use the six-factor solution due to interpretability. The six factor solution explained 76.6% of the variance for Twin 1 whereas a general factor only explained 48.6%. The corresponding figures for Twin 2 were 75.3% and 47.0% respectively. This suggested that a one-factor model was not adequate.	Indeterminate	Adequate
Witting et al. (2008) [174]	CFA	A six-factor model in the Twin 1 group showed a bad fit (chi2(df=137)=789.03, GFI=.924; NFI=.956, RMSEA=.067, AIC 895.08). The results for Twin 2 were similar.	Insufficient	Adequate
Wolpe et al. (2017) [175]	PCA	Two-component structure was found with the first component explaining 76.66% of variance, and the second component explaining 6.16% of variance.	Indeterminate	Adequate
Wylomanski et al. (2014) [176]	EFA	Six-factor structure was found, explaining 71.4% of variance.	Indeterminate	Adequate
Wylomanski et al. (2014) [176]	CFA	Six-factor model did not fit the data. The model was adjusted based on modification indices, adding covariance between error terms of four item pairs: 7–10, 15–16, 3–4 and 8–9. Each pair of items included a similar content. This adjusted model showed a good fit (Q = 2.8, CFI = 0.98, RMSEA = 0.06 and SRMR = 0.03).	Insufficient	Adequate

FSFI: Female Sexual Function Index; PCA: Principal Component Analysis; CFA: Confirmatory Factor Analysis; EFA: Exploratory Factor Analysis

3.2.3 Internal consistency

Thirty-six studies reported on internal consistency of the FSFI-19 [54, 104, 109, 111, 114, 115, 122, 124, 128, 131–133, 137, 140, 141, 143, 145, 150, 151, 154, 156, 159, 161, 163–167, 169, 171, 173, 174, 176, 177, 184, 185] (*Supplementary Table 7.7*). Methodological quality was rated as "very good" [54, 104, 109, 111, 114, 115, 128, 131, 133, 137, 140, 145, 150, 154, 159, 161, 164–167, 169, 171, 173, 174, 176, 184, 185] , "adequate" [141, 151, 163], "doubtful" [124, 132], or "inadequate" [122, 177]. The "inadequate" ratings were due to reporting of Cronbach's Alpha values for only the total score of the FSFI-19. The "doubtful" ratings were due to the reporting of Cronbach's Alpha for an adapted version of a subscale.

Twenty-three studies of "very good" quality [54, 104, 109, 111, 114, 115, 128, 133, 137, 140, 145, 150, 159, 161, 164, 166, 167, 169, 171, 173, 174, 184, 185], two studies of "adequate" quality [151, 163], one study of "doubtful" quality [132], and one study of "inadequate" quality [177] reported Cronbach's Alpha values that were rated as

sufficient (α ≥ .70 and ≤ .95) for all subscales. Four studies of "very good" quality [131, 154, 165, 176], one study of "adequate" quality [141], one study of "doubtful" quality [124], and one study of "inadequate" quality [122] reported multiple Cronbach's Alpha values that were rated as insufficient (α < .70 or > .95).

Four studies reported on internal consistency of the FSFI-6 [55, 179, 180, 182] (*Supplementary Table 7.7*). Methodological quality was rated as "very good" [55, 179, 182], or "inadequate" [180]. The inadequate rating was due to unclear reporting on which items the Cronbach's Alpha was calculated. The evidence of internal consistency was rated as indeterminate for all four studies, as unidimensionality of the FSFI-6 was not investigated (see *Structural validity*), which is a prerequisite for interpreting internal consistency.

3.2.4 Test-retest reliability

Twenty-one studies reported on test-retest reliability of the FSFI-19 [54, 111, 114, 117, 122, 128, 131, 132, 143, 150, 154, 156, 159, 161, 164–166, 171, 175–177] (*Table 3.3*). Methodological quality was rated as "adequate" [117, 128, 150, 175, 176], "doubtful" [54, 111, 114, 122, 131, 132, 143, 154, 156, 159, 161, 164, 171, 177], or "inadequate" [165, 166]. The "doubtful" ratings were due to the use of Pearson Correlation instead of the Intraclass Correlation Coefficients. The "inadequate" ratings were due to a very small sample size ("other flaws" in COSMIN methodological quality) [165], or due to dissimilar test conditions [166].

Five studies of "adequate" quality [117, 128, 150, 175, 176], twelve studies of "doubtful" quality [54, 111, 114, 131, 132, 143, 154, 156, 161, 164, 171, 177], and two studies of "inadequate" quality [165, 166] reported test-retest values that were rated as sufficient. Two studies of "doubtful" quality [122, 159] reported test-retest values that were rated as insufficient.

Two studies reported on test-retest reliability of the FSFI-6 [55,180] (*Table 3.3*). Methodological quality was rated as "adequate" [180], or "doubtful" [55]. The "doubtful" rating was due to use of Pearson Correlation instead of the Intraclass Correlation Coefficient. One study of "doubtful" quality [55] reported test-retest values that were rated as sufficient. One study of "adequate" quality reported test-retest values that were rated as insufficient.

Table 3.3. Test-retest reliability of the FSFI.

Reference	Coefficient	Total score	DE	AR	LU	OR	SA	PA	Rating	Quality
FSFI-19										
Anis et al. (2011) [111]	Correlation	0.98	0.92	0.98	0.97	0.98	0.96	0.97	Sufficient	Doubtful
Bartula et al. (2015) [114]	Correlation		0.86	0.82	0.78	0.8	0.76	0.75	Sufficient	Doubtful
Borello-France et al. (2008) [117]	ICC	0.91	0.84	0.86	0.82	0.9	0.79	0.88	Sufficient	Adequate
Chang et al. (2009) [122]	Correlation	0.69							Insufficient	Doubtful
Fakhri et al. (2012) [128]	ICC	0.77	0.84	0.78	0.86	0.82	0.79	0.73	Sufficient	Adequate
Filocamo et al. (2014) [131]	Correlation	0.95	0.93	0.93	0.95	0.92	0.92	0.93	Sufficient	Doubtful
Ghassamia et al. (2013) [132]	Correlation	0.82	0.66				0.72	0.78	Sufficient	Doubtful
Liu et al. (2016) [143]	Correlation	0.84		0.68	0.83				Sufficient	Doubtful
Nowosielski et al. (2013) [150]	ICC day 7	0.83	0.83	0.89	0.85	0.88	0.87	0.8	Sufficient	Adequate
	ICC day 28	0.75	0.8	0.86	0.8	0.81	0.78	0.73		
Rehman et al. (2015) [154]	ICC	0.99	1	1	0.99	0.98	0.99	1	Sufficient	Doubtful
Rillon-Tabil et al. (2013) [156]	Correlation	0.99							Sufficient	Doubtful
Rosen et al. (2000) [54]	Correlation	0.88	0.83	0.85	0.86	0.8	0.83	0.79	Sufficient	Doubtful
Ryding et al. (2015) [159]	Correlation	.77 - .95	.67 -.89	.62 - .90	.35 - .85	.65 - .86	.65 - .86	.10 - .90	Insufficient	Doubtful
Sidi et al. (2007) [161]	Correlation		0.87	0.77	0.95	0.97	0.95	0.86	Sufficient	Doubtful
Sun et al. (2011) [164]	Correlation	.80 - .86	.72 - .85	.78 - .95	.74 - .93	.85 - .89	.80 - .86	.69 - .90	Sufficient	Doubtful
Takahashi et al. (2011) [165]	ICC		.73 - 1.00	.73 - 1.00	.73 - 1.00	.73 - 1.00	.73 - 1.00	.73 - 1.00	Sufficient	Inadequate
Ter Kuile et al. (2006) [166]	Correlation	0.93	0.72	0.85	0.82	0.71	0.9	0.97	Sufficient	Inadequate
Verit et al. (2007) [171]	Correlation	.90 - .92	.79 - .81	.85 - .87	.85 - .88	.83 - .87	.83 - .85	0.89	Sufficient	Doubtful
Wolpe et al. (2017) [175]	ICC	0.94							Sufficient	Adequate
Wylomanski et al. (2014) [176]	ICC	0.99	0.97	0.99	0.97	0.96	0.89	0.99	Sufficient	Adequate
Zachariou et al. (2017) [177]	Correlation	0.91							Sufficient	Doubtful
FSFI-6										
Isidori et al. (2010) [55]	Correlation	0.95							Sufficient	Doubtful
Lee et al. (2014) [180]	ICC	0.61							Insufficient	Adequate

Reference	Coefficient	Total score	DE	AR	LU	OR	SA	PA	Rating	Quality
FSFI-BC (34 items)										
Bartula et al. (2015b) [183]	Correlation		.72 - .88		.71 - .72	.63 - .85	0.86	.77 - .80	Sufficient	Doubtful

FSFI: Female Sexual Function Index; DE: Desire; AR: Arousal; LU: Lubrication; OR: Orgasm; SA: Satisfaction; PA: Pain

3.2.5 Construct validity (hypothesis testing)

3.2.5.1 Known-group comparison

Twenty-three studies reported on known-group comparison [54, 109, 111, 115, 124, 128, 132, 141, 145, 146, 150, 155, 156, 159, 161, 164–167, 171, 173, 176, 177] of the FSFI-19 (*Supplementary Table 7.8*). Known-group differences were investigated in relation to urological/gynaecological patients versus controls [109, 132, 141, 171], FSD patients versus controls [54, 111, 128, 146, 150, 161, 164, 166, 173], cancer treatment modality [115], patients with hypoactive sexual desire disorder (HSDD) versus controls [124, 145, 155, 159], diabetic patients versus controls [156], premenopausal women versus postmenopausal women [165, 176], age [167], marriage status [176], and women experiencing subjective sexual distress versus controls [177]. Methodological quality was rated as "adequate" for all twenty-three studies. In all twenty-three studies the known-group comparisons provided evidence of sufficient construct validity, as the hypothesized differences between groups were confirmed.

None of the studies that investigated the FSFI-6 [55, 179–182] reported on known-group comparisons.

3.2.5.2 Convergent validity

Forty-nine studies reported on convergent validity of the FSFI-19 (*Supplementary Table 7.9*). The FSFI-19 was compared to measures measuring sexual function and satisfaction [73, 81, 85, 110, 112, 113, 116, 118–120, 123, 124, 126, 128, 139, 146, 148, 150, 152, 155, 158–160, 162, 163, 165, 169, 169, 170, 183, 184], quality of life [114, 183], mental health measures [115, 132], physical functional problems [115, 125, 127, 129, 142, 147, 157, 165, 168, 172, 174, 178], relationship quality [115, 132, 145], and body image [130, 135, 136, 149, 152].

Methodological quality was rated as "adequate" [73, 81, 85, 110, 112– 116, 118–120, 123–127, 129, 130, 132, 135, 136, 139, 142, 145, 147–149, 152, 152, 157, 159, 160, 162, 163, 165, 168–170, 172, 174, 178, 184], "doubtful" [128, 150, 155, 183], or "inadequate" [146]. The "inadequate" rating was due to serious concerns regarding the measurement properties of the comparator instrument. The "doubtful" ratings were

due to concerns regarding the measurement properties of the comparator instrument. Twenty-eight studies of "adequate" rating [73, 81, 110, 114, 118, 119, 123–126, 130, 139, 142, 147–149, 152, 157–160, 162, 163, 165, 168, 169, 172, 178, 184], two studies of "doubtful" rating [128, 155], and one study of "inadequate" quality [146], provided correlations rated as sufficient. Fourteen studies of "adequate" rating [85, 112, 115, 116, 120, 127, 129, 132, 135, 136, 145, 152, 170, 174], and two studies of "doubtful" rating [150, 183], provided correlations rated as insufficient. One study was rated as indeterminate, as not enough information was given for a reliable interpretation [113].

Four studies reported on convergent validity of the FSFI-6 (*Supplementary Table 7.9*). The FSFI-6 was compared to coital frequency [179], educational level [179], partner educational level [179], age [179], partner age [179], waist circumference [179], hot flush intensity [179], FSFI-19 [180], British National Survey of Sexual Attitudes and Lifestyles - Sexual Function [181], Menopause Rating Scale [182], and Hospital Anxiety and Depression Scale [182]. Methodological quality was rated as "adequate" [180–182], or "doubtful" [179]. The "doubtful" rating was due to concerns regarding the measurement properties of the comparator instrument. Two studies of "adequate" quality [180, 181] reported correlations rated as sufficient. One study of "adequate" quality [182], and one study of "doubtful" quality [179] reported correlations rated as insufficient.

3.2.5.3 Divergent validity

Eight studies reported on divergent validity of the FSFI-19 [54, 109, 114, 141, 150, 159, 167, 183] (*Supplementary Table 7.10*). Methodological quality was rated as "adequate" [109, 114, 141, 183], or "doubtful" [54, 150, 159]. The "doubtful" ratings were due to lack of information on the measurement properties of the comparator instrument.

Two studies of "adequate" quality [109, 114], and three studies of "doubtful" quality [54, 159, 167] reported low correlation coefficients that were rated as sufficient. Two studies of "adequate" quality [141, 183], and one study of "doubtful" quality [150] reported multiple correlation coefficients >.30 and were rated as insufficient.

None of the studies that investigated the FSFI-6 [55, 179–182] reported on divergent validity.

3.2.6 Criterion validity

Ten studies reported on criterion validity of the FSFI-19 using the gold standard of FSD or HSDD diagnosis [111, 128, 133, 144, 150, 159, 161, 166, 173, 177] (*Table 3.4*). Methodological quality was rated as "very good" [111, 128, 133, 144, 150, 159, 161, 173], as "adequate" [166], or as "doubtful" [177]. The "doubtful" rating was due to a

small sample size ("other flaws" in COSMIN methodological quality) [177]. One study did not report an Area Under the Curve (AUC) and was thus rated indeterminate [159]. All remaining studies reported AUC values that were rated as sufficient.

Two studies reported on criterion validity of the FSFI-6 using FSD diagnosis as the gold standard [55,180] (*Table 3.4*). Methodological quality was rated as "very good" for both studies. Both studies reported AUC values that were rated as sufficient.

Table 3.4. Criterion validity of the FSFI.

Reference	Instrument	AUC	Cut.off	Sensitivity	Specificity	PPV	PNV	Rating	Quality
FSFI-19									
Anis et al. (2011) [111]	FSD diagnosis	0.99		0.97	0.93			Sufficient	Very good
Fakhri et al. (2012) [128]	FSD diagnosis	0.91		0.82	0.86			Sufficient	Very good
Gerstenberger et al. (2010) [133]	HSDD diagnosis			.70 - .97	.84 - .97			Sufficient	Very good
Ma et al. (2014) [144]	FSD diagnosis Total FSFI		23.45	0.67	0.73			Sufficient	Very good
	Low desire	0.73	< 2.8	0.55	0.78				
	Arousal disorder	0.74	< 3.16	0.62	0.77				
	Lubrication disorder	0.85	< 4.06	0.86	0.7				
	Orgasm disorder	0.85	< 3.9	0.83	0.74				
	Sexual pain	0.79	< 3.9	0.65	0.81				
Nowosielski et al. (2013) [150]	FSD diagnosis	0.93		0.87	0.83	0.86		Sufficient	Very good
Ryding et al. (2015) [159]	HSDD diagnosis			0.96	0.97			Indeterminate	Very good
Sidi et al. (2007) [161]	FSD diagnosis	0.99		0.99	0.97			Sufficient	Very good
Ter Kuile et al. (2006) [166]	FSD diagnosis	0.98		0.95	0.92	0.96	0.89	Sufficient	Adequate
Wiegel et al. (2005) [173]	FSD diagnosis	0.9	26.55	.88 - .89	.71 - .73			Sufficient	Very good
Zachariou et al. (2017) [177]	FSD diagnosis	0.86		0.72	0.93			Sufficient	Doubtful
FSFI-6									
Isidori et al. (2010) [55]	FSD diagnosis	0.98	19	0.96	0.91	0.95	0.93	Sufficient	Very good
Lee et al. (2014) [180]	FSD diagnosis	0.95		0.9	0.86			Sufficient	Very good

FSFI: Female Sexual Function Index

3.2.7 Data synthesis

The synthesized ratings of the measurement properties across all studies can be found in Table 3.5.

Table 3.5. Ratings of measurement properties.

Measurement Property	Rating of Measurement Property	Quality of Evidence
FSFI-19		
Structural Validity	Inconsistent	Low
Internal Consistency	Sufficient	Moderate
Reliability	Sufficient	Low
Measurement Error	Indeterminate	
Construct Validity	Inconsistent	Moderate
Criterion Validity	Sufficient	High
Responsiveness	Indeterminate	
FSFI-6		
Structural Validity	Indeterminate	
Internal Consistency	Indeterminate	
Reliability	Inconsistent	Low
Measurement Error	Indeterminate	
Construct Validity	Inconsistent	Very low
Criterion Validity	Sufficient	Moderate
Responsiveness	Indeterminate	

The evidence of structural validity of the FSFI-19 was rated as inconsistent, because six-factor, five-factor, and other factor structures were reported. The evidence was evaluated as low quality because of this inconsistency, as well as a risk of bias as many studies reported a PCA instead of an EFA or CFA. The evidence of internal consistency of the FSFI-19 was rated as sufficient but of moderate quality, due to 15.8% (N = 6) of studies reporting insufficient internal consistency. The evidence of test-retest reliability of the FSFI-19 was rated as sufficient but of low quality, due to 13.0% (N = 3) of studies reporting insufficient test-retest reliability, as well as risk of bias as many studies reported Pearson Correlation instead of the Intraclass Correlation Coefficient. The use of Pearson Correlations are problematic, as they do not control for systematic error variance, which is a product of measuring the same individual twice. The Intraclass Correlation Coefficient controls for this systematic error variance, and without this control, test-retest reliability may be overestimated [186,187]. The evidence of construct validity (hypothesis testing) of the FSFI-19 was rated as inconsistent with moderate quality, as 28.6% (N = 18) of studies reported insufficient values. The evidence of criterion validity

of the FSFI-19 was rated as sufficient with high quality. The evidence of measurement error and responsiveness of the FSFI-19 were rated as indeterminate, as no data was reported on these measurement properties.

Evidence of structural validity of the FSFI-6 was rated as indeterminate as no data was reported on this measurement property. The evidence of internal consistency of the FSFI-6 was rated as indeterminate, as evidence for unidimensionality was missing. The evidence of test-retest reliability of the FSFI-6 was rated as inconsistent and of low quality, due to risk of bias as only two studies reported on test-retest reliability of which one study was of doubtful methodological quality. The evidence of construct validity (hypothesis testing) of the FSFI-6 was rated as inconsistent with very low quality, as there was as many studies reporting sufficient (50%) as insufficient values (50%), as well as a risk of bias due to methodological quality of the studies. The evidence of criterion validity of the FSFI-6 was rated as sufficient and of moderate quality, as the evidence was based on only two studies. The evidence of measurement error and responsiveness of the FSFI-6 were rated as indeterminate, as no data was reported on these measurement properties.

3.3 Discussion

This systematic review investigated the evidence of the measurement properties of the FSFI-19 [54], and FSFI-6 [55]. Concerning the FSFI-19, the evidence on internal consistency was sufficient and of moderate quality. The evidence on test-retest reliability was also sufficient, but of low quality due to some inconsistencies and many studies not using the ICC. The evidence on criterion validity was also sufficient and of high quality. The evidence on structural validity was inconsistent and of low quality. Studies found either evidence for the theorized six-factor structure, a five-factor structure (with a merging of desire and arousal) or structures with less than five factors. Evidence on construct validity was inconsistent and of moderate quality. No data was found on measurement error and responsiveness.

Concerning the FSFI-6, the evidence on criterion validity was sufficient and of moderate quality. Evidence on reliability was inconsistent with low quality of evidence, due to a high risk of bias. Evidence on construct validity was inconsistent with very low quality of evidence, due to as many studies reporting sufficient as insufficient values as well as a risk of bias due to methodological quality. The evidence of structural validity, internal consistency, measurement error, and responsiveness were rated as indeterminate.

Regarding the structural validity of the FSFI-19, there was more evidence against than in favour of the hypothesized six-factor structure. This is in line with the revisions made

in the DSM-5 to the model of FSD [108]. While only two papers showed direct support for a five-factor structure [128,143], twelve more studies showed indirect support for a five-factor solution through use of PCAs [54,104,115,119,128,132,143,150,154, 173,183,184]. However, other PCAs resulted in other factor structures: four factors [164,173,174], three factors [122,134,138], or two factors [175]. Based on the wide range of reported factor structures, it may be that the factor structure of the FSFI-19 is different for different subgroups or nationalities. In fact, some studies investigated factor structures in subgroups and results suggests there are different factor structures of the FSFI-19 among women with FSD and women without FSD [137,173]. Unfortunately, neither study performed a test of measurement invariance, and as such there is no direct evidence for this hypothesis. Such differences in factor structure in different subgroups may be related to a number of theoretical positions. For one, differing motivations for sex for women with arousal disorder versus without arousal disorder [188], may suggest that arousal and desire may be a singular motivation (i.e. a singular construct) for women without arousal disorder, but not for women with an arousal disorder. Furthermore, the position that FSD represent a spectrum of disorders with extensive overlap [189–191], implies that the constructs measured by the FSFI-19, may be different for women suffering from differing (yet overlapping) sexual disorders.

Nevertheless, based on this systematic review we conclude that the use of the six subscales may not be valid in all patient groups. Instead, we see compelling evidence to merge the subscales of arousal and desire. For confirming whether these subscales should be merged, and whether the constructs measured by the FSFI-19 are different for subpopulations of women, a large-scale validation study focusing on testing measurement invariance across patient subgroups and nationalities, as well as multiple factor structures through use of Confirmatory Factor Analyses is needed. In the meantime we recommend researchers to perform and report on structural validity of the FSFI-19 when presenting the results of their studies, to ensure valid interpretation of their results.

While we rated the evidence on internal consistency as sufficient for the FSFI-19, it needs to be noted that unidimensionality of the subscales is a pre-requisite or interpreting their internal consistency. As the structural validity of the FSFI-19 is shaky at best, our rating of the evidence on internal consistency is mostly intended as an interpretation for the subscales that are found as being unidimensional in most analyses: Lubrication, satisfaction, and pain.

With respect to the structural validity of the FSFI-6, validation studies investigating the unidimensional nature of the instrument are of importance, as no studies investigated this measurement property. The FSFI-6 results in one score representing FSD in general, and it is crucial to determine whether all items represent the construct of FSD

in a unidimensional manner. However, structural validity of the FSFI-6 is likely not so straightforward: as the six best-performing items of the FSFI-19 were selected for each domain, any and all overlap of constructs of the subscales of the FSFI-19 will also be represented in the FSFI-6. Furthermore, as the FSFI-6 is a composite of multiple constructs, it is unlikely to be unidimensional as it is based on a formative model instead of a reflective model. This raises issues with the interpretation of the FSFI-6 total score, as it may not reflect one general construct of FSD. For a total score of a multidimensional instrument, it is unclear what exact construct is represented by the total score.

The evidence on internal consistency of the IIEF-6 can not yet be determined, as the unidimensionality (a prerequisite for internal consistency) has not yet been determined. However, if the FSFI-6 is found to be unidimensional, internal consistency is likely to be rated as sufficient, as three studies of very good quality found values of Cronbach's Alpha that would be rated as sufficient.

Research into measurement error and responsiveness is necessary as well. With the high use of the FSFI-19 and FSFI-6 in clinical practice and clinical research, it is of importance to know which change can be distinguished from measurement error. To further the knowledge, we recommend researchers performing a test-retest reliability studies to calculate the Limits of Agreement [186] or Smallest Detectable Change [192]. Furthermore, an anchor-based study is recommended to determine the Minimal Important Change to be able to interpret the Limits of Agreement or Smallest Detectable Change.

Combining the concerns surrounding structural validity, inconsistent findings on multiple measurement properties, the low quality of many of the included studies, and the missing information on multiple measurement properties; questions are raised on the validity and reliability of the FSFI-19 and FSFI-6 as measures of FSD. The content validity of the FSFI-19 has been challenged previously [193], where the FSFI-19 was described as a measure of vaginal intercourse, and not FSD. Combining concerns regarding content validity, as well as our concerns regarding structural validity, it is unclear whether the FSFI measures FSD, or a selection of symptoms related to FSD. Regardless of these concerns, evidence for criterion validity is strong. Pragmatically, the FSFI is a good screening tool for the current definition of FSD. However, from a psychometric point of view, the above concerns are serious. Given the high frequency of use of both the FSFI-19 and the FSFI-6 in clinical screening for FSD, as well as an outcome measure for clinical trials, it is of importance that more research is performed into the measurement properties and content validity.

A limitation of this review is that we did not investigate content validity. Content validity needs to be established before other measurement properties can be evaluated [38]. A future investigation of content validity is warranted. Another limitation of this review is the use of a precise rather than a sensitive search filter regarding measurement properties. The sensitive filter was developed to capture every relevant hit, at the expense of capturing more false-positive search hits. Meanwhile the specific filter was developed to capture as many relevant hits, while decreasing the number of false-positive search hits. The sensitivity of the precise filter was 93% in a random set of PubMed records, while the sensitivity of the sensitive search filter was 97% [61]. The use of the precise filter was a pragmatic choice over the available sensitive filter as the initial search encompassed 39 PROMs (including the FSFI-19 and FSFI-6), and the sensitive filter would provide too many hits for feasible screening. The possibility remains that the precise filter missed some validation studies of the FSFI-19 and FSFI-6.

3.4 Conclusions

Based on this systematic review, we conclude that with respect to internal consistency, reliability, and criterion validity, the FSFI-19 meets psychometric criteria, but has not been shown to meet psychometric criteria for structural validity, measurement error, construct validity, and responsiveness. Evidence on structural validity suggests a merging of the subscales arousal and desire. Such a merging of subscales has consequences for the interpretation of the FSFI-19 in both clinical practice and research. To investigate this possible adjustment to the FSFI-19, as well as the suggestion that factor structures may be population-dependent; a large-scale cross-cultural study design or an individual patient data meta-analysis, applying Confirmatory Factor Analysis, measurement invariance tests, and calculating the Limits of Agreement and/or Smallest Detectable Change, is recommended.

The FSFI-6 meets psychometric criteria with respect to criterion validity. Structural validity, internal consistency, reliability, measurement error, construct validity, and responsiveness require further research. Most importantly for future research is determining the unidimensionality of the FSFI-6. Regardless of these concerns, evidence for criterion validity is strong for both the FSFI-19 and FSFI-6, and pragmatically, they are good screening tools for the current definition of FSD.


4

Chapter 4

The measurement properties of the EORTC IN-PATSAT32

This chapter was published as Neijenhuijs, K. I., Jansen, F., Aaronson, N. K., Brédart, A., Groenvold, M., Holzner, B., Terwee, C.B., Cuijpers, P., & Verdonck-de Leeuw, I. M. (2018). A systematic review of the measurement properties of the European Organisation for Research and Treatment of Cancer In-patient Satisfaction with Care Questionnaire, the EORTC IN-PATSAT32. *Supportive Care in Cancer, 26(8),* 1-10. doi: 10.1007/s00520-018-4243-9.

Abstract

Purpose: The EORTC IN-PATSAT32 is a patient reported outcome measure (PROM) to assess cancer patients' satisfaction with in-patient health care. The aim of this study was to investigate whether the initial good measurement properties of the IN-PATSAT32 are confirmed in new studies.

Methods: Within the scope of a larger systematic review study (Prospero ID 42017057237), a systematic search was performed of Embase, Medline, PsycINFO, and Web of Science for studies that investigated measurement properties of the IN-PATSAT32 up to July 2017. Study quality was assessed, data were extracted, and synthesized according to the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) methodology.

Results: Nine studies were included in this review. The evidence on reliability and construct validity were rated as sufficient and of the quality of the evidence as moderate. The evidence on structural validity was rated as insufficient and of low quality. The evidence on internal consistency was indeterminate. Measurement error, responsiveness, criterion validity, and cross-cultural validity were not reported in the included studies. Measurement error could be calculated for two studies, and was judged indeterminate.

Conclusion: In summary, the IN-PATSAT32 performs as expected with respect to reliability and construct validity. No firm conclusions can be made yet whether the IN-PATSAT32 also performs as well with respect to structural validity and internal consistency. Further research on these measurement properties of the PROM is therefore needed as well as on measurement error, responsiveness, criterion validity, and cross-cultural validity. For future studies, it is recommended to take the COSMIN methodology into account.

The evaluation of patient health care experiences is relevant for improving health care [194], and as a patient reported outcome measure (PROM) in clinical cancer trials [195]. While multiple PROMs are available to measure patient satisfaction with care [28–32], these PROMs lack international validations [33]. To assess patient satisfaction with health care and to enable cross-cultural comparison of patient health care experiences, the Quality of Life Group of the European Organisation for Research and Treatment of Cancer (EORTC) developed the IN-PATSAT32 [33].

The IN-PATSAT32 is a 32-item PROM assessing hospitalized cancer patients' satisfaction with care. It includes eleven multi-item scales designed to assess: doctors' technical skills (3 items), nurses' technical skills (3 items), doctors' interpersonal skills (3 items), nurses' interpersonal skills (3 items), doctors' information provision (3 items), nurses' information provision (3 items), doctors' availability (2 items), and nurses' availability (2 items), other hospital staff's interpersonal skills and information provision (3 items), waiting time (2 items), and hospital access (2 items). Three single-item scales address the exchange of information, comfort, and general satisfaction.

The initial development and validation study of the IN-PATSAT32 was carried out in 647 patients from eight European countries and Taiwan, and yielded good psychometric results [33]. Multitrait item scaling (MIS) indicated that the structure of the IN-PATSAT32 coincided in most part to the hypothesized structure of items and subscales. Internal consistency and test-retest reliability were satisfactory (α = .80 - .96; ICC = .70 - .85, respectively). Multi- and single-item scales showed evidence of convergent validity when compared to the the Oberst Perception of Care Quality and Satisfaction Scale [29], and divergent validity with the EORTC QLQ-C30 [196]. Finally, validity was supported by the ability of the PROM to distinguish between patients with different levels of intention to recommend the hospital to others [33].

Over a decade after the initial development of the IN-PATSAT32 it is of interest to investigate whether these initial good results regarding the measurement properties of the IN-PATSAT32 are confirmed in other studies, to ensure that it performs as expected in diverse clinical and cultural settings. The aim of the current study was to perform a systematic review of the measurement properties of the IN-PATSAT32, as tested in individual validation studies. Evaluating measurement properties requires weighing many variables on both the level of the study, and on the level of the PROM. Therefore, the current study used the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) criteria for assessing measurement properties of PROMs [38,39,62,63].

4.1 Methods

4.1.1 Literature search strategy

The literature search was part of a larger systematic review (Prospero ID 42017057237 [197]) investigating the validity of 39 different PROMs measuring quality of life of cancer survivors included in an eHealth application called "Oncokompas" (Amsterdam, the Netherlands) [8–11]. The databases Embase, Medline, PsycINFO, and Web of Science were searched for publications that investigated measurement properties of these 39 PROMs including the EORTC IN-PATSAT32. The search terms were the PROM's name, combined with search terms for cancer, and a precise filter for measurement properties [61]. The first search was performed in July 2016. The full search terms can be found in Appendix A. An additional search (up to July 2017) was performed using the same search terms, and a subsequent manual search in Google Scholar and Pubmed for missing records, to search for recent studies.

4.1.2 Inclusion and exclusion criteria

Studies were included that reported original data on cancer patients, and on at least one of the following measurement properties of the IN-PATSAT32 as defined by the COSMIN taxonomy [38,39,62,63]: internal consistency, reliability, measurement error, structural validity, hypothesis testing (for construct validity), criterion validity, cross-cultural validity and responsiveness. Validation studies on other PROMs, that also reported original data on the IN-PATSAT32 were included. Studies that were only available as abstracts or conference proceedings were excluded, as well as non-English publications. Titles and abstracts, and the selected full-texts were reviewed by two independent raters (KN and FJ). Disagreements were discussed until verbal agreement on consensus.

4.1.3 Data extraction

Two independent researchers (KN and FJ) extracted information from eligible papers on each of the measurement properties defined by the COSMIN taxonomy [39]. Relevant data included the type of measurement property, its outcome, and information on methodology. Disagreements were discussed until verbal agreement on consensus.

4.1.4 Data synthesis

Data synthesis consisted of three steps. First, the quality of the methodology of the included studies was rated using the 4-point scoring system of the COSMIN checklist [39,62,63]. Methodological aspects regarding design requirements and preferred statistical methods, specific to each measurement property under consideration were rated as either "poor", "fair", "good", or "excellent". The methodological quality was operationalized per measurement property per study as the lowest score they received on any of the methodological aspects. The final ratings can be found in Appendix B.

Second, criteria for good measurement properties were applied to the results of the included studies, following the COSMIN guidelines for systematic reviews of PROMs [38]. Each measurement property in each individual study was rated as sufficient (+), insufficient (–) or indeterminate (?), according to predefined criteria. For indeterminate ratings, the methodological rating was non-applicable. All of these ratings were qualitatively summarized to determine the overall rating of the measurement property. If all studies indicated a sufficient, insufficient, or indeterminate rating for a specific measurement property, the overall rating of this measurement property was accordingly. If there were inconsistencies between studies, explanations were explored (e.g. differences in methodological quality). If explanations were found, they were discussed until consensus was reached and taken into account during interpretation. If no explanations were found, the overall rating would be inconsistent (±).

Third, we used the modified GRADE approach [38] to rate the quality of the evidence available for the measurement properties of the IN-PATSAT32. This approach takes into account (i) methodological quality, (ii) directness of evidence, (iii) inconsistency of results, and (iv) precision of evidence. The overall quality of evidence was rated as high, moderate, low, or very low. Measurement properties that were rated as indeterminate in the previous step, did not receive a rating as there was no evidence to rate. All ratings (methodological quality, measurement property rating, and GRADE rating) were rated by one researcher (KN), whose ratings were checked by a second independent researcher (AH). Discrepancies in ratings were discussed until verbal agreement on consensus.

4.2 Results

4.2.1 Search results

The initial search identified 980 abstracts of which 10 were relevant to the IN-PATSAT32 (*Figure 4.1*). Three abstracts and one full-text were excluded for not providing unique information on a measurement property. One study not captured by the search, but known to the authors was added before data-extraction. The search update up to July 2017 identified three more abstracts of which one was excluded for not providing unique information on a measurement property. No full-texts were excluded from this search update. In total nine studies were included in this review (See *Supplementary Table 7.11*). These nine studies reported on the structural validity (6 studies), internal consistency (5 studies), reliability (2 studies), and hypothesis testing (6 studies) of the IN-PATSAT32, but lacked information on measurement error, criterion validity, responsiveness, and cross-cultural validity. We were able to calculate measurement error for two studies.

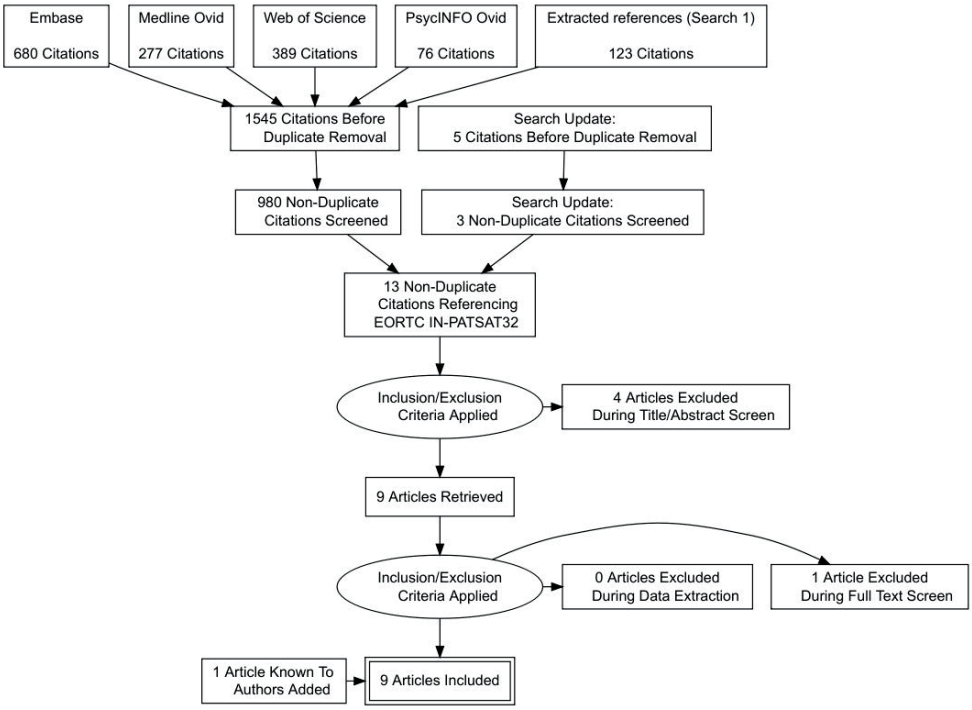

Figure 4.1. PRISMA diagram.

4.2.2 Structural validity

Six studies reported on structural validity. Methodological quality of these studies was rated as "good" [198], "fair" [199], or "poor" [200–203] (*Table 4.1.*). The poor ratings were due to using Multitrait Item Scaling (MIS) instead of confirmatory or exploratory factor analysis (CFA / EFA). The fair score was due to lack of information about the handling of missing values. Results of the MIS analyses were consistent across studies, as well as with the original validation study. However, MIS is an indirect way of testing structural validity. Therefore, no conclusions can be drawn on basis of these studies. Two articles [198,199] presented results of Principal Component Analyses (PCA). Hjörleifsdóttir [198], of "good" quality, extracted four components with an eigenvalue > 1, with a balanced distribution of explained variance. Pishkuhi [199], of "fair" quality, extracted five components with an eigenvalue > 1, and one of those components explained most of the variance. The factor structures found in these two studies were inconsistent with the 11 subscale (and 3 single-item scale) model as reported in the initial study [33], leading to an insufficient rating.

Table 4.1. Structural validity of the IN-PATSAT32.

Reference	Methodology	Outcome	Rating Structural validity	Quality
Arraras et al., 2009 [200]	Multitrait Item Scaling	Most items exceeded correlations of .4 with other items in their own scale, except for items 29 and 30 (Hospital Access). Most items had a higher correlation with other items in their own scale than items in other scales, except for items 14 (Nurse Interpersonal Skills), 21, 22 (Nurse Availability), 24 (Other Staff Interpersonal Skills), and 30 (Hospital Access).	Indeterminate	Poor
Hjörleifsdóttir et al., 2010 [198]	Mulitrait Item Scaling	All items exceeded correlations of .4 with other items in their own scale. The weakest scale was 'satisfaction with service and care organisation', in which 50% of the items correlated higher with other items in their own scale than other items in other scales. The strongest scale was 'satisfaction with nurses' conduct', in which 92% of items correlated higher with other items in their own scale than other items in other scales.	Indeterminate	n/a
Hjörleifsdóttir et al., 2010 [198]	Principal Component Analysis	Four components were extracted with an eigenvalue >1, explaining 67.4% of variance. The components can be identified as: Satisfaction with nurses (24.7% variance), satisfaction with doctors (21% variance), satisfaction with information (13.6% variance), and satisfaction with service (8% variance).	Insufficient	Good
Obtel et al., 2017 [203]	Multitrait Item Scaling	All items exceeded correlations of .4 with other items in their own scale. All items had higher correlations with other items in their own scale than items in other scales.	Indeterminate	Poor
Pishkuhi et al., 2014 [199]	Mulitrait Item Scaling	All items exceeded correlations of .8 with other items in their own scale. All items had higher correlations with other items in their own scale than items in other scales.	Indeterminate	n/a
Pishkuhi et al., 2014 [199]	Principal Component Analysis	Five components were extracted with an eigenvalue > 1, explaining 71.1% of variance. The components can be identified as: Satisfaction with nurses (45.4% variance), satisfaction with services & care organization (9.5% variance), satisfaction with doctors (8.1% variance), satisfaction with doctors' information provision (4.7% variance), and satisfaction with nurses' information provision (3.2% variance).	Insufficient	Fair
Zhang et al., 2014 [201]	Mulitrait Item Scaling	All items exceeded correlations of .4 with other items in their own scale. Fifty percent of items had a higher correlation with other items in their own scale than items in other scales.	Indeterminate	Poor
Zhang et al., 2015 [202]	Mulitrait Item Scaling	All items exceeded correlations of .4 with other items in their own scale. Six out of twenty-nine items had a significantly lower correlation with items in their own scale than items in other scales.	Indeterminate	Poor

4.2.3 Internal consistency

Five studies reported on internal consistency of the IN-PATSAT32, and their methodological quality was rated as "good" [198], "fair" [199], or "poor" [200–203]. The main reason for the poor ratings was that the unidimensionality of the scales was not tested appropriately. The values for Cronbach's alpha of five studies [199–203] are presented in Supplementary Table 7.12. One other study [198] presented Cronbach's alpha values for scales they had established: nurse satisfaction (α = .95), doctor satisfaction (α = .93), information satisfaction (α = .91), and service satisfaction (α = .67). However, as these scales do not represent the subscales recommended for this questionnaire [33], this study is not included in Supplementary Table 7.12, nor further taken into account. All but one subscale (Hospital Access) showed Cronbach's alpha values that would qualify for a sufficient rating. However, as none of the studies provided any evidence of unidimensionality for the subscales, Cronbach's alpha cannot be properly interpreted [204]. The inconsistency of Cronbach's alpha coefficients across studies is noteworthy for the subscale Hospital Access (α = .36 - .86).

4.2.4 Reliability

Two studies [199,203] reported on test-retest reliability (See *Supplementary Table 7.13*). Methodological qualities were rated as "fair" due to lack of information about the handling of missing values [199,203], not reporting the type of correlation coefficient [199], and a short time interval (30 minutes) [203]. One study [199] showed high test-retest correlations (r > .85), leading to a sufficient rating on test-retest reliability. However, as the type of correlation coefficient was not reported, it is unclear whether these values represent appropriate estimates of test-retest reliability [186,187]. The other study [203] showed acceptable test-retest correlations (ICC > .70), except for Doctors' Availability (ICC = .64) and General Comfort (ICC = .67), leading to a sufficient rating.

4.2.5 Measurement error

While none of the studies presented results regarding measurement error, the Standard Error of Measurement (SEM) and Smallest Detectable Change (SDC) could be calculated for the two studies reporting test-retest reliability [199,203]. Methodological quality was "good", due to the need to calculate measurement error indirectly (*Table 4.2*). Since no Minimal Important Change (MIC) was reported, a criterion for good measurement error could not be applied. While there is no evidence for or against good measurement error, the SDC could be compared to the maximum range of the subscales. The SDC represents the minimum change score over time of which we can be certain does not represent measurement error. Most SDC scores were between 20-30, representing 20-30% on the 100-point scale. There were a few notable outliers: Doctor Availability (29.17 - 46.40), Waiting Time (25.05 - 44.70), and Hospital Access (29.39 - 34.48).

Table 4.2. Measurement error (Standard Error of Measurement & Smallest Detectable Change) of the IN-PATSAT32.

Reference	DrTech	DrInt	DrInfo	DrAva	NTech	NInt	NInfo	NAva	SInt	WT	HA	IE	HC	OA	Rating	Quality
Pishkuhi et al., 2014 [199]															?	Fair
SEM	9.53	10.82	7.99	10.53	7.88	6.47	6.05	8.67	6.8	9.03	10.6					
SDC	26.42	29.98	22.13	29.17	21.85	17.93	16.77	24.58	18.85	25.05	29.39					
Obtel et al., 2017 [203]															?	Fair
SEM	7.83	8.09	8.38	16.81	7.84	12.08	8.8	9.1	10.92	16.13	12.44	10.7	14.4	14.88		
SDC	21.69	22.41	23.23	46.6	21.74	33.49	24.38	25.22	30.26	44.7	34.48	29.66	39.93	41.24		

SEM = Standard Error of Measurement; SDC = Smallest Detectable Change; DrTech = Doctor Technical Skills; DrInt = Doctor Interpersonal Skills; DrInfo = Doctor Information Provision; DrAva = Doctor Availability; NTech = Nurse Technical Skills; NInt = Nurse Interpersonal Skills; NInfo = Nurse Information Provision; NAva = Nurse Availability; SInt = Other Staff Interpersonal Skills; WT = Wait Times; HA = Hospital Access; IE = Information Exchange; HC = Hospital Comfort; OA = Overall Satisfaction; ? = Indeterminate

4.2.6 Construct validity (hypothesis testing)

4.2.6.1 Known-group comparison

Three studies performed known-group comparison, a comparison between groups that are known to show differences on the measured construct. Known group differences were investigated with respect to age [201], educational level [201], tumour stage [202], time since diagnosis [202], and satisfaction with care [200]. The methodological quality of these studies was rated as "fair" [200] or "poor" [201,202]. The poor scores were due to not providing a priori hypotheses, while the fair score was due to lack of information about the handling of missing values (*Table 4.3*). The known-group comparisons investigated by Arraras [200] were based on a priori hypotheses, and provide sufficient evidence of construct validity. Due to not providing a priori hypotheses, the results of Zhang [201,202] were rated as indeterminate.

Table 4.3. Known-group validity of the IN-PATSAT32.

Reference	Comparison groups	Outcome	Rating	Quality
Arraras et al., 2009 [200]	Low vs. high score on the Oberst perception of care quality and satisfaction scale	Significant differences in all IN-PATSAT32 areas except nurse availability. Patients with higher Oberst scores had greater care satisfaction.	Sufficient	Fair
Arraras et al., 2009 [200]	Low vs. high score on item investigating intention to recommend the hospital or ward to others	Significant differences in all IN-PATSAT32 areas except nurse availability. Patients with higher intention to recommend the hospital or ward had greater care satisfaction.	Sufficient	Fair
Zhang et al., 2014 [201]	Patients < 58 years vs. patients = 58 years	Patients < 58 years scored significantly higher than patients = 58 years, except on nurse availability and hospital comfort	Indeterminate	Poor
Zhang et al., 2014 [201]	Patients who finished lower than compulsory education vs. patients who finished compulsory or higher education	Patients who had finished compulsory education scored significantly higher than patients who had not finished compulsory education.	Indeterminate	Poor
Zhang et al., 2015 [202]	Patients who finished lower than compulsory education vs. patients who finished compulsory or higher education	Patients who had finished compulsory education scored significantly higher on technical skills, interpersonal skills, information provision, and availability of both doctors and nurses. Effect sizes were small (< .50) in for all scales.	Indeterminate	Poor
Zhang et al., 2015 [202]	Patients with metastatic vs. non-metastatic tumors	Patients with metastatic tumors scored significantly higher on nurses' conduct, other hospital staffs' interpersonal skills information provision scales. Effect sizes were small (< .50) except for nurses' interpersonal skills (-.55), nurses' information provision (-.57), and nurses' availability (-.51).	Indeterminate	Poor

Zhang et al., 2015 [202]	Patients with > 2 months diagnostic time vs. patients with < 2 months diagnostic time	Patients with > 2 months diagnostic time scored significantly higher on nurses' conduct, other hospital staffs' interpersonal skills information provision scales. Effect sizes were small (< .5) except for nurses' technical skills (-.55), and nurses' interpersonal skills (-.50).	Indeterminate	Poor

4.2.6.2 Convergent validity

Four studies reported on convergent validity, and compared the IN-PATSAT32 to the EORTC QLQ-INFO25 (measuring patient perceptions of information received and their information needs) [205], the Oberst patients' perception (measuring the quality of care received and how well the care meets patients' expectations [29]) [200], and the EORTC QLQ-C15-PAL (measuring quality of life of patients with incurable cancer [207]) [208]. The methodological quality of these studies was rated as either "good" [205], "fair" [200,206], or "poor" [208]. The poor score was due to not providing a priori hypotheses [208]. The fair scores were due to lack of information about the handling of missing values [206], or due to lack of information about a priori hypotheses [200] (*Table 4.4*). Two studies [200,206], of "fair" quality, demonstrated moderate correlations (r > .40) with related constructs, indicative of sufficient convergent validity. Asadi-lari [205], of "good" quality, and Aboshaiqah [208], of "poor" quality, found low correlations (r < . 40) for most of the constructs that were hypothesized to be related to the IN-PATSAT32, indicating insufficient convergent validity.

Table 4.4. Convergent validity of the IN-PATSAT32.

Reference	Comparison instrument	Correlations	Rating	Quality
Aboshaiqah et al., 2016 [208]	EORTC QLQ-C15-PAL	IN-PATSAT32 general satisfaction correlated with physical function (r = .21), emotional function (r = .32), and global health status (r = .26).	Insufficient	Poor
Arraras et al., 2009 [200]	Oberst patients' perception of care quality and satisfaction scale	Oberst medical care scale correlated with the IN-PATSAT32 doctor scales (.62 - .71). The Oberst information adequacy scale correlated with the IN-PATSAT32 doctor information provision (.70) and nurses' information provision (.62) scales. The Oberst quality of nursing scale correlated with the IN-PATSAT32 nurse scales (.60 - .69). The Oberst self-care information scale correlated with doctors' (.60) and nurses' (.61) information provision.	Sufficient	Fair
Arraras et al., 2010 [206]	EORTC QLQ-INFO25	Doctors' information provision (.61), nurses' information provision (.46), other staff interpersonal skills (.42) correlated with the QLQ-INFO25 item regarding information satisfaction. Single items regarding information provision of the IN-PATSAT32 correlated with QLQ-INFO25 items measuring similar constructs (.30 - .61), with more similar theoretical items correlating higher (> .40).	Sufficient	Fair

Asadi-lari et al., 2015 [205]	EORTC QLQ-INFO25	Doctors' information provision (.23), nurses' information provision (.39), and other staff interpersonal skills (.20) correlated with the QLQ-INFO25 item regarding information satisfaction. Single items regarding information provision of the IN-PATSAT32 correlated with the QLQ-INFO25 items measuring similar constructs (.15 - .41).	Insufficient	Good

4.2.6.3 Divergent validity

Four studies reported on divergent validity, and compared the IN-PATSAT32 scales to scales of the EORTC QLQ-C30 (measuring health related quality of life in cancer patients [196]. Their methodological quality was rated as "fair" [199,200] or "poor" [201,202]. The poor scores were due to not providing a priori hypotheses. The fair score of Arraras [200] was due to the lack of detail in formulated a priori hypotheses, while the fair score of Pishkuhi [199] was due to lack of information about the handling of missing values. One study of "fair" quality found no significant correlations [199], and one study of "fair" quality [200] and two studies of "poor" quality [201,202] found correlations smaller than .40, indicative of sufficient divergent validity.

4.2.7 Data Synthesis

The synthesized ratings of the measurement properties can be found in Table 4.5. Internal consistency was rated indeterminate as no tests of unidimensionality were reported. Measurement error was rated indeterminate as no MIC was reported and could not be calculated with the available data. Structural validity was rated insufficient with evidence of low quality. Test-retest reliability and construct validity (hypothesis testing) were judged to be sufficient, both with evidence of moderate quality. The indeterminate findings [201,202] for construct validity were not taken into account in this synthesis, as they did not provide evidence for or against construct validity. Studies of "poor" quality were outweighed by studies with better quality. One study of "good" quality provided insufficient evidence on convergent validity for construct validity [205], while three studies of "fair" quality provided sufficient evidence on known-groups comparison and convergent validity for construct validity [199,200,206].

Table 4.5. Ratings of measurement properties.

Measurement Property	Rating of Measurement Property	Quality of Evidence
Structural Validity	Insufficient	Low
Internal Consistency<U+2060>	Indeterminate	
Reliability	Sufficient	Moderate
Measurement Error	Indeterminate	
Construct Validity	Sufficient	Moderate

4.3 Discussion

This systematic review investigated the current evidence up to July 2017 regarding the measurement properties of the EORTC IN-PATSAT32 [33]. Nine studies were included in this review. The evidence on reliability and construct validity were rated as sufficient and of moderate quality evidence. The evidence on structural validity was rated as insufficient and of low quality. The evidence on internal consistency was indeterminate, as the assumption of unidimensionality was not investigated. Measurement error, responsiveness, criterion validity, and cross-cultural validity were not reported in the studies reviewed.

With respect to structural validity, the developers of the IN-PATSAT32 postulated an a priori scale structure, and provided support for that structure in their original validation study [33]. In the studies that reported on structural validity [198–203] MIS or PCA was applied instead of CFA. The findings of the PCA analyses [198,199] are of particular interest as they revealed fewer scales compared to the original 11-scale (and 3 separate single-item) factor structure [33].

Future studies investigating structural validity may inform their theorized factor structures based on these results. They may consider performing CFAs to test the posited 11-scale structure, but also two factor structures which seem plausible, given the results of the reported PCAs [198,199]:

1. A first-order factor structure where the relevant items load on one of four factors: (i) satisfaction with nurses; (ii) satisfaction with doctors; (iii) satisfaction with services & care; and (iv) information provision;

2. A second-order factor structure where all items load on the originally developed scales. The originally developed scales will then load on the relevant second-order factors: (i) satisfaction with nurses; (ii) satisfaction with doctors; (iii) satisfaction with services & care; and (iv) information provision.

Test-retest reliability was rated as sufficient in the present review although of moderate quality evidence. When this property is examined in future studies, it is important that the Intraclass Correlation Coefficient is used to control for systematic error variance. Without controlling for systematic error variance, test-retest reliability may be overestimated [186,187].

In the present review, none of the studies reported on *measurement error*. We calculated the standard error of measurement (SEM) and smallest detectable change (SDC) based

on the data of two studies. Relating the SDC to the maximum range of the scale, showed that most values were around 20-30% of the scales, although a number of outliers were observed. To interpret these data, information on the minimal important change (MIC) is needed. This should preferably be derived from anchor-based methods. Subsequently, the MIC should be compared to the *measurement error* to determine if the scales can detect small but important changes that are not an artefact of *measurement error*.

Cross-cultural validity was explored in the original validation process [33]. In future studies, this can be investigated further by performing measurement invariance tests for subsamples in CFAs, or by pooling data of multiple international studies to perform measurement invariance tests for language. Unfortunately, it is not possible to assess *criterion validity*, as there is no "gold standard" for assessing patient satisfaction. *Responsiveness* could be investigated through longitudinal studies of changes in patient satisfaction with care.

A limitation of this review is the use of a precise rather than a sensitive search filter regarding measurement properties. The sensitivity of the precise filter was 93% in a random set of PubMed records, while the sensitivity of the sensitive search filter was 97% [61]. The use of the precise filter was a pragmatic choice over the available sensitive filter as the initial search encompassed 39 PROMs (including the IN-PATSAT32), and the sensitive filter would provide too many hits for feasible screening. Although we also performed a manual search and found no missing records, the possibility remains that the precise filter missed validation studies of the IN-PATSAT32. Furthermore, because we included only papers published in English, we may have missed information from studies published in other languages.

Based on this systematic review, we conclude that with respect to test-retest reliability and construct validity, the IN-PATSAT32 performs as expected in diverse clinical and cultural settings. However, no firm conclusions can be made as to whether the IN-PATSAT32 performs as well with respect to structural validity and internal consistency. Further research on these measurement properties of the EORTC IN-PATSAT32 is therefore needed as well as on measurement error, responsiveness, criterion validity, and cross-cultural validity. For future studies, it is recommended to take the COSMIN methodology into account.

Intermezzo

Reflections on Measurement Error

This chapter is based on the proceedings of the "Measurement Error in Psychological Science" session at the 2019 meeting of the Society for the Improvement of Psychological Science. A more complete write-down of this intermezzo can be found on my Github[1] [209].

1 https://kneijenhuijs.github.io/SIPS-2019-Measurement-Error/

Abstract

Background: Measurement Error represents the minimum amount of change measured by a measurement tool, of which we can be sure is not an artefact of systematic error. In a large-scale systematic review, we found that 4.14% of validation articles reported on measurement error, and measurement error could be calculated for another 3.82% of articles. To illustrate the implications measurement error has on clinical research, a simulation study was conducted.

Methods: Simulations were run on a hypothetical randomized controlled trial for the treatment of depression as measured by the BDI-II. Baseline values and a decrease over time for untreated depression (control condition) were extracted from literature. The Minimal Clinically Important Difference (MCID) was used as a measure of effect size for the further decrease over time of the treatment condition. Three parameters were systematically varied across simulations: sample size (250 / 500 / 750), effect size (0*MCID / 1*MCID / 2*MCID / 3*MCID), and measurement error (0% / 10% / 20% / 30% / 40%). Each parameter combination was simulated 5000 times.

Results: The relative bias is the bias of the coefficient of interest. The relative bias became more biased from near zero (with no measurement error) to -0.5 (with 30% and 40% measurement error). Furthermore, effect sizes showed more relative bias. ETA Squared is a measure of effect size. The ETA Squared ranges from 0 to 0.525 when there is 0% measurement error, dependent on the effect size parameter. Every ETA squared drifted further towards zero with more added measurement error.

Conclusions: The results of the simulation showed an increase in bias with the addition of more measurement error. Furthermore, this effect seemed to be stronger for higher effect sizes. The result of this bias is a decrease of effect size, which is especially dramatic upwards of 20% measurement error. It appears that measurement error affects power to detect a true effect.

In Chapter 2-4 I discussed the systematic review of the measurement properties of three PRMs used in Oncokompas. As discussed in the introduction of this dissertation, these are not the only PRMs we investigated. A report discussing the full results of this systematic review are published elsewhere [51]. After data extraction of 314 validation articles of all these PRMs, we found a surprising lack of reports on measurement error. A total of 13 validation articles (4.14%) reported on measurement error, of which 9 reported a Standard Error of Measurement, 3 reported Limits of Agreement, and 2 reported Person Standard Error. For a total of 12 validation articles (3.82%) we could calculate Standard Error of Measurement and Smallest Detectable Change. As such, only 25 (7.96%) of all articles reported data relevant to Measurement Error. This lack of reported Measurement Error was an inspiration to chair a session called "Measurement Error in Psychological Science" at the annual meeting of the Society for the Improvement of Psychological Science in 2019. What follows is the information presented at this session, as well as the discussion following this presentation.

Measurement Error represents the minimum amount of change measured by a measurement tool, of which we can be sure is not an artefact of systematic error. The amount of Measurement Error is ideally smaller than the Minimal Clinically Important Difference, which represents the minimum amount of change measured by a measurement tool, which is judged to be represent a clinically meaningful change for the patient. We want the Measurement Error to be smaller, so that we can be sure that when we measure a clinically meaningful change, it is not an error.

Measurement Error may have large implications for both clinical practice (e.g. "did the patient improve / deteriorate over time?") as well as research settings (e.g. "is the change caused by our intervention large enough to warrant implementation?"). To illustrate the latter, I drafted a simulation to represent a research setting in clinical psychology.

I.I Methods

I simulated a RCT for the treatment of depression, with depression being measured by the BDI-II. Why the BDI-II? Because I could find some nifty statistics on it to make the simulation more rooted in reality.

The BDI-II has a range of 0 - 63, which can be categorized into four categories:

1. 0-13: minimal depression

2. 14-19: mild depression

3. 20-28: moderate depression

4. 29-63: severe depression

A systematic review from 2013 provided me with data on baseline means and standard deviations of clinical samples [210]. I pooled these means and standard deviations, which resulted in a pooled baseline mean of 24.1 and pooled baseline sd of 11.4. These numbers will be used to represent our fixed intercept in the data generation.

A meta-analysis found that the score on the BDI-II decreased by 15.7% for untreated depression groups [211], which translates to a mean decrease of 9.828. This decreases will represent our fixed slope across time.

We are assuming that our treatment group has a further decrease in depression. Because the Minimal Clinically Important Difference (MCID) is related to measurement error, and because it has been studied in the BDI-II, I am using the MCID as an 'effect size'. The MCID of the BDI-II is 18% [212], leading to a MCID of 4.338. We're going to use multiples of the MCID as a parameter in the data generation. Because the group has already decreased to 24.1 - 9.828 = 14.272, we can test out effect sizes up to 3 times the MCID, after which the score on the BDI-II would be very close to zero.

Measurement error was added based on a percentage of the range of the BDI-II. In my experience (and that of some colleague psychometricians I asked, #AnecdotalEvidence) a Smallest Detectable Change representing 20% of the range of the measurement instrument is a regular finding. As such, I decided to use measurement error ranging from 0% up to 40%.

A detailed rundown of the code of the simulation can be found on my Github[2] [209].

I.2 Results

Five measures were used to investigate the bias introduced by measurement error: Relative bias, mean absolute bias, mean standard error, ETA squared, and Empirical Detection Rates. In this intermezzo I only present the relative bias and ETA squared. For a discussion of the remaining measures, see my Github[2] [209].

I.2.1 Relative bias

The relative bias is the bias of the coefficient of interest. The bias is divided by the value of the coefficient used in the data generation. We expect that with more measurement error, this bias becomes larger.

2 https://kneijenhuijs.github.io/SIPS-2019-Measurement-Error/

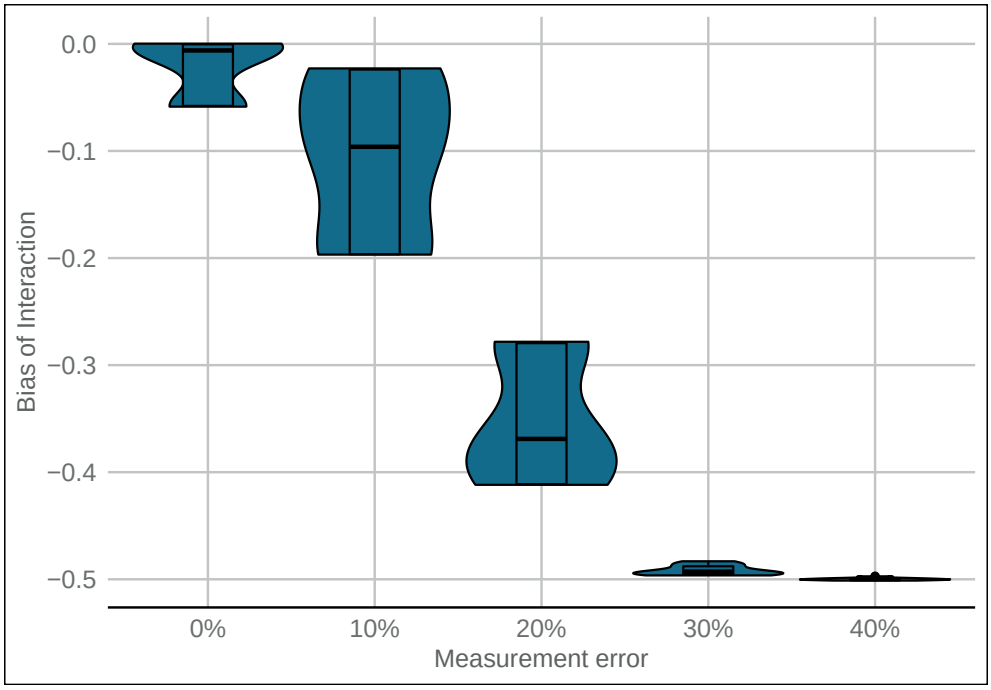

Figure I.1. Relative Bias of Interaction.

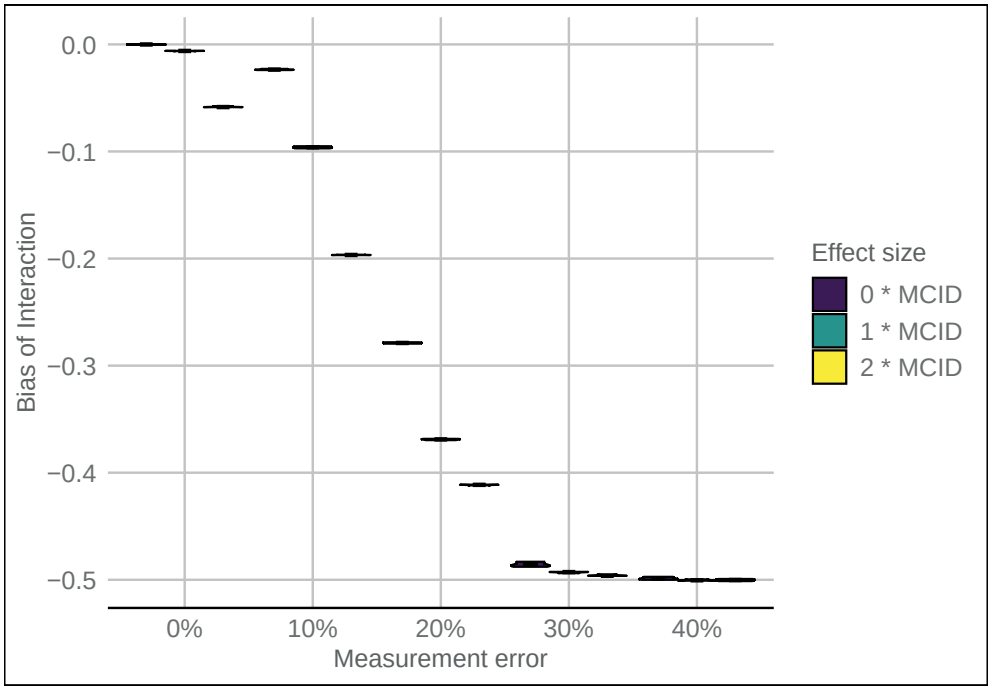

Figure I.2. Relative Bias of Interaction By Effect Size.

As you can see, our interaction coefficient becomes biased from near zero (with no measurement error) to -0.5 (with 30% and 40% measurement error). A clear and also very predictable result: More measurement error creates more bias in our coefficient. How does this relate to effect size?

There is very little variance within the violin plots, making it somewhat difficult to read. However, we can see that higher effect sizes seem to show more bias in the interaction coefficient.

I.2.2 ETA Squared

ETA Squared is a measure of effect size. We expect the ETA Squared to deviate away from the "expected" ETA Squared given how large the coefficient is. I'm not a 100% sure how large the ETA squared should be given the size of the coefficient, so our 0% measurement error serves as our comparison unit. While the plot including the sizes of the coefficient is more informative, I do want to show the plot without them first.

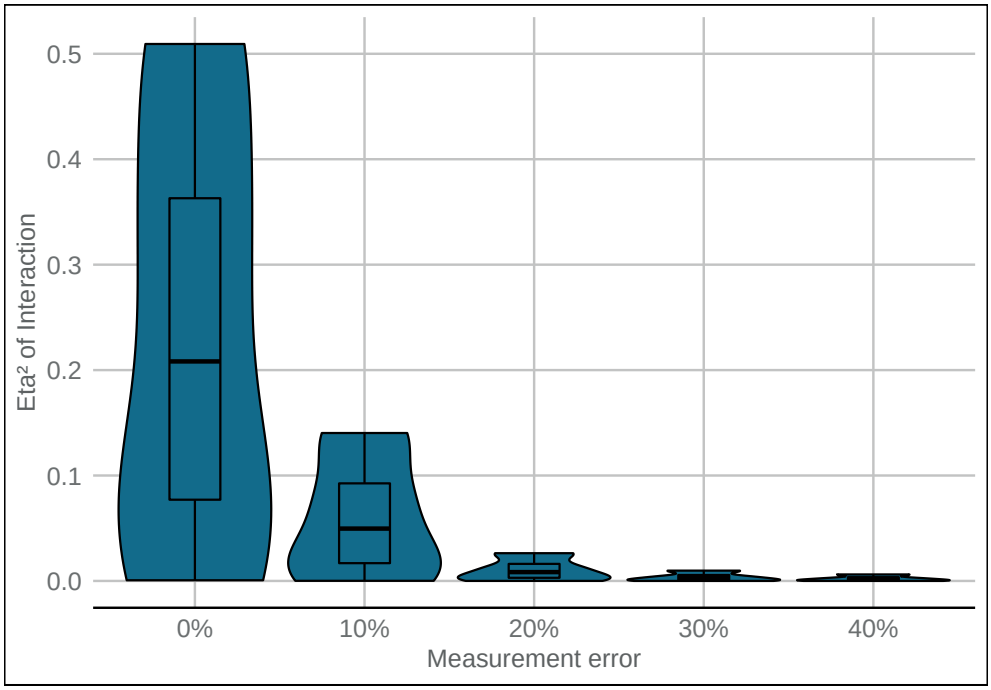

Figure I.3. Eta Squared of Interaction.

With 0% measurement error, we see a nice distribution of ETA Squared, which makes sense as this is an aggregate of all the simulations regardless of coefficient size. The more measurement error we add, the further ETA Squared deviates towards zero.

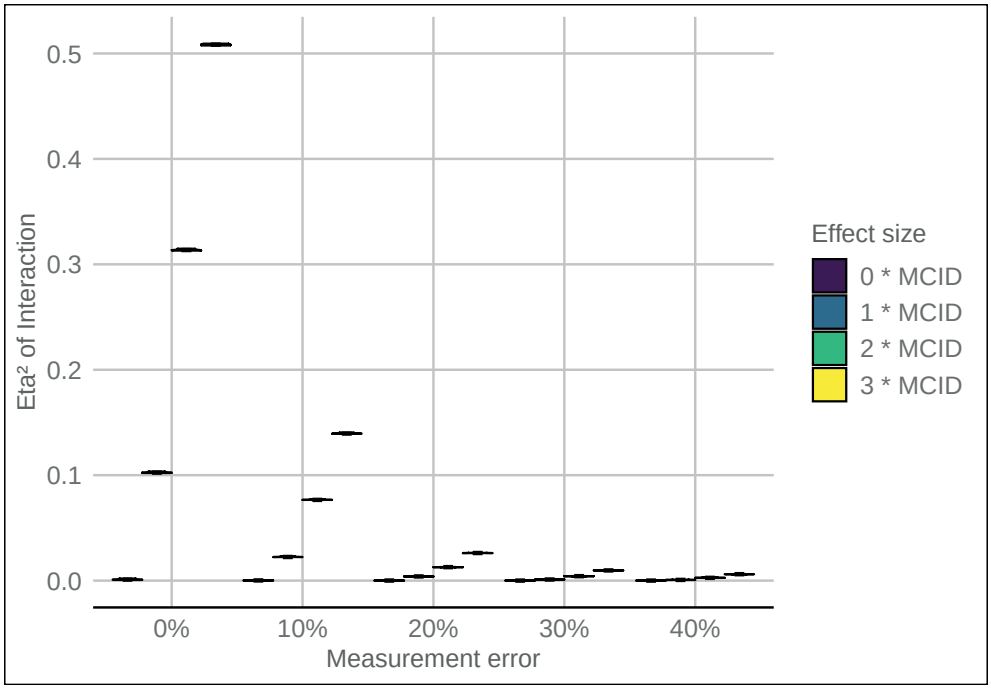

Figure I.4. Eta Squared By Effect Size.

Unfortunately there is very little variance within the violin plots, making this plot hard to read. The sizes of the coefficient go from 0 to 13 from left to right. So we can see that the ETA Squared ranges from 0 to 0.525 when there is 0% measurement error. Every ETA squared drifts further towards zero with more added measurement error.

I.3 Simulation limitations

There are three limitations I have found relevant to this simulation. All of the limitations are related to how the data is generated. For a detailed view of the parameters I mention below, see the code on on my Github[3] [209].

1. A lot of parameters in the data generation are super-arbitrary. In particular, the calculation of the variance-covariance matrix for generation of the measurement error has multiple parameters that can be improved upon. The intra-individual correlation can be improved by finding any justification for a certain value. While the variance of the measurement error has some justification, it is all based on personal logic, instead of literature. The parameters in the variance-covariance matrix for generation of the random effects are equally (if not more) arbitrary.

3 https://kneijenhuijs.github.io/SIPS-2019-Measurement-Error/

2. The analysis of the data sets based on zero measurement error result in a lot of singular models (i.e. models where the random effects are close to zero). This might actually be the explanation for why I found no trend in Empirical Detection Rates. I tried adding some more random error to the data generation, but this created an unreasonably high amount of data points that equal zero on the BDI-II, which is not realistic.
3. On the subject of error, the data generation definitely needs some as random effects and measurement error are definitely not the only sources of error in data. Like I stated, I tried adding some more random error to the data generation, but this created an unreasonably high amount of data points that equal zero on the BDI-II, which is not realistic.

I.4 Simulation conclusion

The results of the simulation unequivocally show an increase in bias with the addition of more measurement error. Furthermore, this effect seems stronger for higher effect sizes. The result of this bias is a decrease of effect size, which is especially dramatic upwards of 20% measurement error. It seems measurement error affects power to detect a true effect.

I.5 Discussion

Our session did not end with me presenting this simulation. With a group of over 30 researchers, we discussed several aspects of measurement error and how they relate to our research. The rest of this chapter summarises the results of this discussion. The discussion focused mostly on ways to correct for measurement error.

The room mentioned the use of the Spearman's correction for attenuation, which was developed to correct for measurement error [213]. However, this method has proven to be less than ideal, due to imprecision. A recent pre-print showed how mixed-effect models outperform Spearman's correction through estimation of random effects [214].

One suggestion was offered for measurement properties to be discussed during peer review. What measures were used? What is known in the literature on the measurement properties of these measures? What measurement properties did you find in your sample? This last question is particularly important, as the generalization of measurement properties is an assumption that does not hold true in practice. I feel like this suggestion combines well with the more wide suggestion that peer review should include at least one statistician who is familiar with the analysis methods used. Something which has been mentioned as early as 2006 [215], but unfortunately has not yet seen wide implementation.

My favourite suggestion from the room: How about a database of measurement properties, to which anyone can upload their data as well as the measurement properties they found in that data? This is the holy grail, but I'm sceptical towards the feasibility. Another project at SIPS is the scienceverse of Lisa DeBruine and Daniël Lakens[4]. They aim to create easy-to-use tools for researchers to create a machine-readable summary of their project. In a project such as this, measurement properties could be a part of the generated summary. Furthermore, through integration with open data databases (e.g. OSF), large datasets could be formed (or found) on which independent researchers could run validity/reliability analyses. The large issue is the need for standardization and centralization of practices, which is a large roadblock to overcome.

I made one suggestion which believe is feasible, but was met with fair scepticism from the room. The suggestion is to have researchers run validation/reliability analyses on datasets that they have already collected. While I made the suggestion that certain analyses can be performed on most datasets and are relatively easy to perform, the room was not convinced. The issue raised was that for validation analyses a large number of expert decisions are necessary. While I agree with this notion, there are certain analyses that do not have to be complicated for them to add to our knowledge. For example, if you have a questionnaire with three subscales, a Confirmatory Factor Analysis can be used to test whether those subscales fit to the data, using accessible R packages, and these days even using JASP[5]. I agree, that when you have a bad fit to the data, and want to explore alternative factor structures, this becomes complicated very quickly. In the future, guidelines or specialized centres to help with these analyses may make this suggestion a feasibility.

I.6 Conclusion

The session "Measurement Error in Psychological Science" at the annual meeting of the Society for the Improvement of Psychological Science in 2019 was successful in obtaining multiple points of view on Measurement Error across disciplines of psychological science. Suggestions and questions raised during the discussion may form a basis for further research into the issue, and more importantly may be a starting point for projects to develop guidelines and tools to help researchers improve knowledge on Measurement Error as well as other measurement properties of their measurement instruments.

4 https://docs.google.com/document/d/1DKhnypsG__XG9k_16smU3IJDYGgnxFP5LHw4P6Qh50g/
5 https://jasp-stats.org/2018/07/03/how-to-perform-structural-equation-modeling-in-jasp/


5

Chapter 5

Validation of Dutch version of eHealth Impact Questionnaire

This chapter was published as Neijenhuijs K. I, van der Hout A., Veldhuijzen E., Scholten-Peeters G. G. M., van Uden-Kraan C. F., Cuijpers P., & Verdonck-de Leeuw I. M (2019). Translation of the eHealth Impact Questionnaire for a Population of Dutch Electronic Health Users: Validation Study. *Journal of Medical Internet Research,* 21(8):e13408. doi: 10.2196/13408.

Abstract

Background: The eHealth Impact Questionnaire (eHIQ) provides a standardized method to measure attitudes of electronic health (eHealth) users towards eHealth. It has previously been validated in a population of eHealth users in the United Kingdom, and consists of 2 parts and 5 subscales. Part 1 measures attitudes toward eHealth in general and consists of the subscales *Attitudes towards online health information* (5 items), and *Attitudes towards sharing health experiences online* (6 items). Part 2 measures the attitude towards a particular eHealth application and consists of the subscales *Confidence and identification* (9 items), *Information and presentation* (8 items), and *Understand and motivation* (9 items).

Objective: This study aimed to translate and validate the eHIQ in a Dutch population of eHealth users.

Methods: The eHIQ was translated and validated in accordance with the COnsensus-based Standards for the selection of health status Measurement INstruments criteria. The validation comprised 3 study samples with a total of 1287 participants. Structural validity was assessed using confirmatory factor analyses and exploratory factor analyses (EFAs;all 3 samples). Internal consistency was assessed using hierarchical omega (all 3 samples). Test-retest reliability was assessed after 2 weeks, using two-way intraclass correlation coefficients (sample 1). Measurement error was assessed by calculating the smallest detectable change (sample 1). Convergent and divergent validity were assessed using correlations with the remaining measures (all 3 samples). A graded response model was fit and item information curves were plotted to describe the information provided by items across item trait levels (all 3 samples).

Results: The original factor structure showed a bad fit in all 3 study samples. EFAs showed a good fit for a modified factor structure in the first study sample. This factor structure was subsequently tested in sample 2 and 3, and showed acceptable to good fits. Internal consistency, test-retest reliability, convergent validity, and divergent validity were acceptable to good for both the original as the modified factor structure, except for test-retest reliability of one of the original subscales, and the 2 derivative subscales in the modified factor structure. The graded response model showed that some items underperformed in both the original and modified factor structure.

Conclusions: The Dutch version of the eHIQ (eHIQ-NL) shows a different factor structure compared with the original English version. Part 1 of the eHIQ-NL consists of 3 subscales: *attitudes towards online health information* (5 items), *comfort with sharing health experiences online* (3 items), and *usefulness of sharing health experiences online* (3 items). Part 2 of the eHIQ-NL consists of three subscales: *motivation and confidence to act* (10 items), *information and presentation* (13 items), and *identification* (3 items).

Currently, patients and care providers are encouraged to use electronic health (eHealth) applications to improve health care including selfmanagement [216,217]. A standardized measure to evaluate eHealth applications throughout the development process is needed. In the Netherlands, more than 98% of the population has access to the internet [218] and the use of eHealth applications is stimulated by both government and health care organizations. Internationally, the access to the internet is also growing rapidly. A standardized measure to evaluate eHealth applications is therefore much needed. However, evaluating eHealth applications is difficult because of a number of factors, including the difficulty of creating controlled experiments and confounding variables such as proficiency with the internet [34], and the continued development of eHealth application in comparison to more traditional forms of health care. Currently, evaluation of eHealth applications usually consists of two components: testing efficacy using randomized controlled trials, and in-depth evaluation of the content of the application using structured and unstructured interviews. These methods require a large investment of time and resources. Given the rapid development of technology, this creates a state of "playing catch-up" for eHealth developers. A standardized way of evaluating eHealth applications can be invaluable in the process of constant development and evaluation. Although some such standardized measures exist (e.g. the System Usability Scale, which measures the usability of software applications), they do not offer similar insight into the user experience as through interviews.

In 2013, Kelly et al. [35] developed the eHealth Impact Questionnaire (eHIQ) to measure the self-reported impact of eHealth on its users. On the basis of 5 themes, which were identified from interviews, the questionnaire consists of 2 parts. The first part (11 items) measures the overall attitude of eHealth users regarding eHealth, consisting of 2 subscales: *attitudes towards online health information* (5 items), and *attitudes towards sharing health experiences online* (6 items). The second part (26 items) measures the attitude of eHealth users regarding a specific eHealth application, consisting of 3 subscales: *confidence and identification* (9 items), *information and presentation* (8 items), *understand and motivation* (9 items). This questionnaire was validated in 2015 for the British eHealth users [36].

The goal of this study was to translate and validate the eHIQ in a Dutch population of eHealth users — resulting in the Dutch version called eHIQ-NL — according to the COnsensus-based Standards for the selection of health status Measurement INstruments (COSMIN) criteria [62]. These criteria provide a systematic roadmap for appropriate analyses and interpretation of different types of validity and reliability. To our knowledge, the eHIQ has not been previously translated and/or validated outside of the original development and validation [35,36].

In the first study (the main study), Dutch users of the website Kanker.nl (an eHealth website for Dutch cancer patients) completed both parts of the eHIQ twice. In the second study, the first part of the eHIQ was completed by Dutch cancer survivors who were invited to participate in a survey on supportive cancer care, which was part of a randomized controlled trial to evaluate the efficacy of Oncokompas (an eHealth self-management application that supports Dutch cancer survivors in finding and obtaining optimal supportive care)[8]. In the third study, the second part of the eHIQ was completed by Dutch patients who had undergone orthopaedic surgery and were participants in a pilot study of an app providing health information regarding pre- and post-operative care.

5.1 Methods

5.1.1 Translation

The questionnaire was translated from English into Dutch by 2 independent translators; 1 eHealth expert and 1 language expert who is a Dutch native and fluent in English. These translations were combined into a single Dutch questionnaire by 2 independent reviewers. In case of discrepancies the final translation was decided by consensus. The Dutch translation was then translated back into English by 2 independent experts language who are English natives and fluent in Dutch. The back-translated questionnaire was compared with the original English version by 2 independent reviewers. Discrepancies between the back-translated and the original English questionnaire were discussed, and some final changes were made. An example copy of the final translated questionnaire can be found in Appendix F.

5.1.2 Recruitment and Procedure

Because of the results of the main study (study sample 1), the eHIQ was subsequently presented to 2 other samples of (prospective) eHealth users (study samples 2 and 3).

5.1.2.1 Study Sample 1

Dutch users of the national website Kanker.nl (an eHealth website for cancer patients) who had signed up to participate in scientific research, were asked to fill in both parts of the eHIQ-NL twice, with an interval of 2 weeks. On the second measurement, they were also asked to answer 2 questions designed to gauge attitudes to eHealth applications; 1 question asked them to grade their satisfaction with Kanker.nl on a 10-point scale (Overall satisfaction), while the other question asked how likely they were to recommend Kanker.nl to a fellow cancer patient (the Net Promoter Score (NPS)). They were further asked to fill in the 5-level EuroQol-5D version (EQ-5D-5L), which measures self-reported health-related quality of life [219].

5.1.2.2 Study Sample 2

A random sample of cancer survivors (breast cancer, colorectal cancer, head and neck cancer or lymphoma) was drawn from the Netherlands Cancer Registry and invited to complete a survey on supportive cancer care, which was part of an RCT investigating the efficacy of Oncokompas (an eHealth self-management application that supports Dutch cancer survivors in finding and obtaining optimal supportive care) [8]. Patients were excluded who had severe cognitive impairment, insufficient mastery of the Dutch language, physical inability to complete a questionnaire, or received palliative care. Participants with internet access filled in the first part of the eHIQ-NL during the survey on supportive care. They were also asked to fill in the Functional, Communicative and Critical Health Literacy (FFCHL) scales (Cronbach's α was .94 in the current sample) - which measures the capacity of individuals to access, understand, and use health information - [220], and the European Organisation for Research and Treatment of Cancer core quality of life questionnaire, version 3.0 (Cronbach's α was .98 in the current sample), which measures cancer-related quality of life [196]. Medical ethical approval was provided by the Medical Ethics Review Board of the VU Medical Center in Amsterdam, the Netherlands (reference number 2015.523).

5.1.2.3 Study Sample 3

Patients were recruited from a single clinic (ViaSana, Mill, The Netherlands) to participate in a pilot study of an app providing health information regarding pre- and post-operative care. Patients were eligible when aged older than 18 years, and had undergone orthopaedic surgery. Patients were excluded if they were not accessible by e-mail. Participants filled in the second part of the eHIQ-NL up to 2 weeks after using the application. Participants also filled in the System Usability Scale (SUS; Cronbach's α was .90 in the current sample), which measures the usability of software applications [221], and 2 questions designed to gauge attitudes to eHealth applications; 1 question asked them to grade the application on a 10-point scale, whereas the other question asked how likely they were to recommend the application to a fellow patient. Medical ethical approval was provided by Medical Ethics Review Board of the Elisabeth Hospital in Tilburg, the Netherlands (reference number METC-T2012-11).

5.1.3 Statistical Analysis

All analyses were performed in R version 3.3.3 [222]. Measurement properties were assessed in accordance with the COSMIN criteria [62].

5.1.3.1 Study Sample 1

First, structural validity was assessed with a combination of confirmatory Factor analyses (CFAs) and exploratory factor analyses (EFAs). All CFAs were run using the cfa function

of the lavaan package [223], whereas all EFAs were run using the efaUnrotate function of the semTools package [224], and Oblimin rotation was applied using the oblqRotate function of the semTools package [224].

Second, internal consistency was assessed by calculating hierarchical omega [225] using the reliability function of the semTools package [224]. Third, test-retest reliability was assessed by calculating a 2-way intraclass correlation coefficient (ICCs) between the 2 measurement times, using the icc function of the irr package [226]. Fourth, measurement error was assessed by calculating the standard error of measurement using the SE.Meas function of the psychometric package [227]. The smallest detectable change was calculated by hand using the standard error of measurement.

Fifth, convergent validity and divergent validity were tested by correlating the subscales of the eHIQ-NL with the questions concerning satisfaction with Kanker.nl and the NPS (a positive correlation was hypothesized), and the EQ-5D-5L of which the items for *Daily activities and Anxiety / Depression* were assumed to show a positive correlation. No correlation was hypothesized to exist between the eHIQ-NL and the remaining EQ-5D-5L items. Correlations were calculated using the rcorr function of the Hmisc package [228].

Sixth and last, a graded response model was fit using the grm function of the ltm package [229]. Item information curves were plotted for each subscale to describe the information provided by items across the item trait level (i.e. the construct measured by the subscale).

5.1.3.2 Study Sample 2

Structural validity was assessed with a combination of CFAs and EFAs. Internal consistency was assessed with hierarchical omega. Divergent validity was tested by correlating the subscales of the eHIQ-NL with the FCCHL, as no correlation was hypothesized to exist. Finally, a graded response model was fit. All analyses were performed using the same functions and R packages as in study sample 1.

5.1.3.3 Study Sample 3

Structural validity was assessed with a combination of CFAs and EFAs. Internal consistency was assessed with hierarchical omega. Convergent validity was tested by correlating the subscales of the eHIQ-NL with the SUS, and the questions concerning the grade of and likelihood of recommending the application, as positive correlations were hypothesized. Finally, a graded response model was fit. All analyses were performed using the same functions and R packages as in study sample 1.

5.2 Results

5.2.1 Study Population

Table 5.1 shows the demographic and clinical characteristics of the 3 study samples. In study sample 1, 304 cancer survivors participated with a mean age of 58.12 years (standard deviation, SD = 11.26) and 177 were female (58.2% (177 / 304)). The study sample consisted of more than 17 cancer diagnoses; most were diagnosed with breast cancer (27% (82 / 340) or prostate cancer (13.8% (42 / 340). The feasibility of the eHIQ-NL was good: of the 304 participants who started the first measurement, 288 (94.7% (288 / 304)) completed the eHIQ-NL. A total of 242 (79.6% (242 / 304)) participants started the second measurement, of which 217 (71.4% (217 / 304)) completed all questionnaires.

In study sample 2, 566 cancer survivors completed the first part of the eHIQ-NL with a mean age of 64.18 years (SD = 10.65) and 351 (62.1% (351 / 565)) were female. The study sample consisted of 4 cancer diagnoses; breast cancer (39.2% (222/566)), colorectal cancer (29.7% (168 / 566)), head and neck cancer (19.1% (108 / 566)) and lymphoma (12.0% (68 / 566)).

In study sample 3, 526 orthopaedic patients completed the second part of the eHIQ-NL with a median age of 59.00 years (interquartile range = 50 - 66), and 267 were female (51% (267 / 526)). The study sample consisted of patients who underwent various orthopaedic surgeries; the main group had undergone a total knee artroplasty (31.1% (164 / 526)).

Table 5.1. Study population: descriptive statistics.

	Age			Gender		Diagnosis		
	N	**Mean**	**SD**	**Gender**	**N**	**Type**	**N**	**%**
Study Sample 1								
	288	58.12	11.26	Male	111	Breast cancer	82	26.97
				Female	177	Miscellaneous cancer	47	15.46
						Prostate cancer	42	13.82
						Lymphoma	20	6.58
						Colon cancer	18	5.92
						Skin cancer	18	5.92
						Lung cancer	17	5.59
						Bladder & Kidney cancer	17	5.59
						Rectal cancer	14	4.61
						Head & Neck cancer	12	3.95
						Esophageal cancer	10	3.29
						Leukemia	10	3.29

	Age			Gender		Diagnosis		
	N	Mean	SD	Gender	N	Type	N	%
						Other	33	8.88
Study Sample 2								
	566	64.18	10.65	Male	214	Breast cancer	222	39.2
				Female	351	Colorectal cancer	168	29.7
						Head & Neck cancer	108	19.1
						Lymphoma	68	12
Study Sample 3	**Median**	**IQR**						
	526	59	50-66	Male	259	Total knee arthroplasty	164	31.2
				Female	267	Total hip arthroplasty	89	16.9
						Anterior cruciate ligament reconstruction	56	10.6
						Knee arthroscopy	47	8.9
						Cuff repair	30	5.7
						High tibial osteotomy	23	4.4
						Lumbar discectomy	17	3.2
						Acromionplasty	14	2.7
						Remaining group	86	16.3

Rest group = shoulder arthroplasty, femoral osteotomy, patella stabilisation (MPFL), mortons neurom, hallux valgus/rigidus, exostosis, talocrual arthrodesis

5.2.2 Study Sample 1

5.2.2.1 Structural Validity

A CFA was run on a 2-level hierarchical model, with the specified subscales as first-order factors, and the 2 different sections (general attitude and specific attitude) as second-order factors. This model had a bad fit (minimum discrepancy per degree of freedom [CMIN] = 2.61, adjusted goodness-of-fit index [AGFI] = .719, comparative fit index [CFI] = .752, Tucker-Lewis index [TLI] = .753, standardized root mean square residual [SRMR] = .076, and root mean square error of approximation [RMSEA] = .075 [.070 - .079]). Inspecting the modification indices revealed cross-loadings of items on the second-order factors. Such cross-loadings made sense when looking at the content of the items (e.g., items on information on the specific eHealth tool showing cross-loadings with general attitude towards health information); however, shifting items from 1 section to another made no theoretical or practical sense. Therefore, 2 CFAs were run separately for each section, removing the second-order factor from the analysis.

The fit for the first part of the questionnaire was better than the first model fit, but not yet acceptable (CMIN = 5.14, AGFI = .796, CFI = .847, TLI = .804, SRMR = .074, and RMSEA = .118 [.103 - .134]). A 3-factor EFA using Oblimin rotation was

run to investigate an alternative to the original factor structure. This model showed a good fit (CMIN = 3.16, AGFI = .989, CFI = .954, TLI = .898, SRMR = .032, and RMSEA = .085 [.065 - .107]). The 3 factors were interpretable (*Table 5.2A*), with the subscale *attitudes towards sharing health experiences online* being split into the two factors *comfort with sharing health experiences online* and *usefulness of sharing health experiences online*. The third factor was identical to the original factor of *attitudes towards online health information*.

The fit for the second part of the questionnaire was also better than the first model fit, but not yet acceptable (CMIN = 3.20, AGFI = .747, CFI = .755, TLI = .731, SRMR = .082, and RMSEA = .087 [.081 = .094]). A 4-factor EFA using Oblimin rotation was run to investigate an alternative to the original factor structure. The model showed a good fit (CMIN = 2.01, AGFI = .988, CFI = .914, TLI = .876, SRMR = .037, and RMSEA = .059 [.051 - .067]), but the factor structure was not clearly interpretable, many items had double loadings, and the fourth factor had very low factor loadings. A 5-factor EFA using Oblimin rotation showed a similar fit (CMIN = 1.93, AGFI = .988, CFI = .928, TLI = .886, SRMR = .033, and RMSEA = .057 [.048 - .065]). While the double loadings were mostly taken care of, the loadings on the fourth and fifth factor were very low.

A 3-factor EFA using Oblimin rotation was run to investigate problematic items. Items 10, 8, 16, 4, 17, and 11 showed double loadings and no clear distinction to any one factor. Removing these items and performing a CFA on the original factor structure resulted in a bad fit (CMIN = 3.39, AGFI = .779, CFI = .786, TLI = .757, SRMR = .084, and RMSEA = .091 [.083 = .099]). Running an EFA using Oblimin rotation on the same subset of items resulted in a good fit (CMIN = 2.22, AGFI = .990, CFI = .920, TLI = .886, SRMR = .041, and RMSEA = .062 [.052 - .072]), but with a different factor structure than theorized (*Table 5.2B*): the first factor being a combination of items from the subscales *confidence and identification and understanding and motivation*, and interpretable as *motivation and confidence to act*; the second factor being identical to the original subscale *information and presentation* with the addition of item 2; and the third factor consisting of three items from the subscale *confidence and identification*, and interpretable as *identification*. The results of these 3-factor EFAs of the first and second part of the eHIQ will henceforth be refered to as the "modified factor structure".

Table 5.2. Structural Validity: Exploratory Factor Analysis factor loadings.

	Item	Factor 1	Factor 2	Factor 3
A: Study Sample 1 - eHIQ-NL Part 1				
	Part 1 - item 9	0.772*		
	Part 1 - item 8	0.663*		
	Part 1 - item 11	0.592*		
	Part 1 - item 6		0.863*	
	Part 1 - item 7		0.759*	
	Part1 - item 10		0.612*	
	Part 1 - item 4			-0.826*
	Part 1 - item 5			-0.670*
	Part 1 - item 3			-0.620*
	Part 1 - item 1			-0.455*
	Part 1 - item 2		0.312*	-0.434*
B: Study Sample 1 - eHIQ-NL Part 2				
	Part 2 - item 23	0.787*		
	Part 2 - item 22	0.676*		
	Part 2 - item 7	0.657*		
	Part 2 - item 21	0.616*		
	Part 2 - item 20	0.592*		
	Part 2 - item 18	0.589*		
	Part 2 - item 1	0.560*		
	Part 2 - item 13	0.401*		
	Part 2 - item 9		0.722*	
	Part 2 - item 6		0.676*	
	Part 2 - item 3		0.541*	
	Part 2 - item 5		0.519*	
	Part 2 - item 26		0.498*	
	Part 2 - item 12		0.481*	
	Part 2 - item 24		0.435*	
	Part 2 - item 2		0.416*	
	Part 2 - item 25		-0.390*	
	Part 2 - item 15			0.842*
	Part 2 - item 14			0.799*
	Part 2 - item 19			0.592*
C: Study Sample 2 - eHIQ-NL Part 1				
	Part 1 - item 9	-0.453*		
	Part 1 - item 11	-0.432	0.351	
	Part 1 - item 8	-0.334	0.333	
	Part 1 - item 7		0.911*	
	Part 1 - item 6		0.791*	
	Part 1 - item 10	-0.371	0.658*	

	Item	Factor 1	Factor 2	Factor 3
	Part 1 - item 4			-0.759*
	Part 1 - item 3			-0.736*
	Part 1 - item 2		0.31	-0.617**
	Part 1 - item 1			-0.596**
	Part 1 - item 5			-0.544*
D: Study Sample 3 - eHIQ-NL Part 2				
	Part 2 - item 22	0.745*		
	Part 2 - item 23	0.669*		
	Part 2 - item 7	0.517*		
	Part 2 - item 1	0.505*	-0.319	
	Part 2 - item 8	0.505*		
	Part 2 - item 10	0.455*	-0.389	
	Part 2 - item 18	0.415*		
	Part 2 - item 13	0.355*		
	Part 2 - item 21	0.204*		
	Part 2 - item 9		-0.849*	
	Part 2 - item 6		-0.847*	
	Part 2 - item 12		-0.693*	
	Part 2 - item 5		-0.631*	
	Part 2 - item 17		-0.613*	
	Part 2 - item 26		0.583*	
	Part 2 - item 2		-0.598*	
	Part 2 - item 25		0.583*	
	Part 2 - item 11		-0.495*	
	Part 2 - item 3		0.447*	
	Part 2 - item 4	0.392	-0.431*	
	Part 2 - item 24		-0.420*	
	Part 2 - item 16		-0.402*	
	Part 2 - item 19			0.916*
	Part 2 - item 15			0.790*
	Part 2 - item 14			0.639*
	Part 2 - item 20	0.463*		0.466*

Standardized Factor Loadings. Loadings < .30 suppressed.

5.2.2.2 Internal Consistency

Table 5.3A and 5.3B shows the results on internal consistency of the original factor structure and the modified factor structure, respectively. All values were acceptable (omega > .70), and the values of the original first part and the modified first part were comparable. The values of the modified second part were better than of the original second part.

Table 5.3. Internal consistency.

	Original factor structure	Modified factor structure	Modified factor structure problematic items				
	Subscale			**Omega**	**Subscale**	**Omega**	**Cronbach's Alpha**
A: Study Sample 1 - eHIQ-NL Part 1							
	Part 1: General attitude			0.84	Part 1: General attitude	0.9	
	Attitudes towards online health information			0.79	Attitudes towards online health information	0.79	
	Attitudes towards sharing health experiences online			0.78	Comfort with sharing health experiences online	0.73	
					Usefulness of sharing health experiences online	0.83	
B: Study Sample 1 - eHIQ-NL Part 2							
	Part 2: Specific attitude			0.9	Part 2: Specific attitude	0.89	
	Confidence and identification			0.85	Motivation and confidence to act	0.85	
	Information and presentation			0.7	Information and presentation	0.78	
	Understanding and motivation			0.81	Identification	0.82	
C: Study Sample 2 - eHIQ-NL Part 1							
	Part 1: General attitude			0.9	Part 1: General attitude	0.91	
	Attitudes towards online health information			0.81	Attitudes towards online health information	0.81	
	Attitudes towards sharing health experiences online			0.88	Comfort with sharing health experiences online	0.76	

	Original factor structure	Modified factor structure	Modified factor structure problematic items			
				Usefulness of sharing health experiences online	0.86	
D: Study Sample 3 - eHIQ-NL Part 2						
	Part 2: Specific attitude		0.87	Part 2: Specific attitude	0.89	0.91
	Confidence and identification		0.92	Motivation and confidence to act	0.85	0.91
	Information and presentation		0.65	Information and presentation	0.7	0.91
	Understanding and motivation		0.83	Identification	0.86	0.91

Table 5.4. Test-retest reliability.

Original factor structure			Modified factor structure		
Subscale	**ICC**	**CI**	**Subscale**	**ICC**	**CI**
Attitudes towards online health information	0.71	0.64 - 0.77	Attitudes towards online health information	0.71	0.64 - 0.77
Attitudes towards sharing health experiences online	0.63	0.54 - 0.7	Comfort with sharing health experiences online	0.62	0.53 - 0.69
Confidence and identification	0.73	0.66 - 0.78	Usefulness of sharing health experiences online	0.53	0.43 - 0.62
Information and presentation	0.72	0.64 - 0.78	Motivation and confidence to act	0.76	0.7 - 0.81
Understanding and motivation	0.74	0.67 - 0.8	Information and presentation	0.73	0.66 - 0.79
			Identification	0.7	0.62 - 0.76

ICC = Intraclass Correlation Coefficient; CI = 95

Table 5.5. Measurement error.

Original factor structure			Modified factor structure		
Subscale	**SEM**	**SDC**	**Subscale**	**SEM**	**SDC**
Attitudes towards online health information	9.14	25.32	Attitudes towards online health information	9.14	25.32
Attitudes towards sharing health experiences online	9.44	26.18	Comfort with sharing health experiences online	12.56	34.81
Confidence and identification	6.79	18.83	Usefulness of sharing health experiences online	10.43	28.9
Information and presentation	5.69	15.77	Motivation and confidence to act	7.19	19.93
Understanding and motivation	6.33	17.54	Information and presentation	5.43	15.05
			Identification	8.94	24.78

SEM = Standard Error of Measurement; SDC = Smallest Detectable Change

5.2.2.3 Test-Retest Reliability

Table 5.4 shows the results on test-retest reliability of the original factor structure and the modified factor structure. All original subscales, except for *attitudes towards sharing health experiences online* (ICC = .63) showed acceptable ICCs (ICC > .70). All modified subscales, except for *comfort with sharing health experiences online* (ICC = .62) and *usefulness of sharing health experiences online* (ICC = .53) showed acceptable ICCs (ICC > .70). The ICCs for the original factor structure and the modified factor structure were comparable.

5.2.2.4 Measurement Error

Table 5.5 shows the results of the measurement error of the original factor structure and the modified factor structure. For the original factor structure, the SDC ranged between 15.77 and 26.18, which represents a measurement error of 15% - 26% of the 100 subscale range. Consequently, we can be 95% certain that a change score larger than 15% to 26% of the subscale range is not an artefact of measurement error. For the modified factor structure the SDC ranged between 15.05 and 34.81, which represents a measurement error of 15% to 35% of the 100 subscale range. The highest SDCs were reported for the Part 1 *attitudes towards sharing health experiences online* (34.81) and *comfort with sharing health experiences online* (28.91) subscales. This makes sense, as both subscales only consisted of 3 items, and small scales are susceptible to high measurement error.

5.2.2.5 Convergent and Divergent Validity

All subscales correlated significantly with both the overall satisfaction and the NPS. The correlations between the subscales of the first part of the eHIQ-NL and the overall satisfaction and the NPS were small (r < .30). There were either no significant or very small (r < .20) correlations with the EQ-5D questions on daily activities, and anxiety and depression. The 3 remaining EQ-5D items did not correlate significantly with any of the eHIQ-NL subscales (*Table 5.6A* and *5.6B*).

5.2.2.6 Graded Response Model

Figure 5.1 shows the item information curves for the original subscales. A number of items of part 1 did not provide much extra information to the subscale: items 1, 2, 8, 9, and 11. Notably, most items in the subscale *attitudes towards sharing health experiences online* provided information at the same item trait levels. A number of items of part 2 also did not provide much extra information to the subscale: items 2, 10, 11, 13, 16, 23, and 25. Notably, items 10, 11, and 16 were items that fit poorly in the factor analysis.

Table 5.6. Convergent and divergent validity.

	Original factor structure					Modified factor structure					
	OHI	**SHEO**	**C&I**	**I&P**	**U&M**	**OHI**	**CSHEO**	**USHEO**	**M&CA**	**I&P**	**ID**
A: Study Sample 1 - Convergent											
Grade	0.17*	0.28***	0.50***	0.47***	0.49***	0.17*	0.23***	0.27***	0.46***	0.49***	0.35***
NPS	0.21**	0.25***	0.48***	0.44***	0.44***	0.21**	0.19**	0.26***	0.43***	0.45***	0.37***
EQ-5D - Daily activities	-0.01	0.07	-0.05	0	-0.05	-0.01	0.08	0.04	-0.05	0	0.01
EQ-5D - Anxiety/Depression	-0.07	0	-0.13	-0.18**	-0.14*	-0.07	0.06	-0.07	-0.15*	-0.18**	-0.02
B: Study Sample 1 - Divergent											
EQ-5D - Mobility	-0.04	0.02	-0.04	-0.05	-0.02	-0.04	0.01	0.02	-0.01	-0.03	0.01
EQ-5D - Selfcare	0.02	0.07	-0.04	-0.03	0.03	0.02	0.06	0.07	0.01	-0.02	-0.04
EQ-5D - Pain	-0.04	0	-0.08	-0.06	-0.08	-0.04	0.03	-0.04	-0.09	-0.06	-0.03
C: Study Sample 2 - Divergent											
FCCHL	0.14**	0.03				0.14**	-0.01	0.06			
EORTC QLQ-C30	0.01	0.01				0.01	0	0.02			
D: Study Sample 3 - Convergent											
System Usability Scale			0.29***	0.53***	0.39***				0.30***	0.55***	0.12**
NPS			0.46***	0.42***	0.52***				0.48***	0.45***	0.31***
Grade			0.59***	0.53***	0.62***				0.58***	0.55***	0.43***

* = <.05, ** = <.01, *** = <.001 OHI = "Attitudes towards online health information"; SHEO = "Attitudes towards sharing health experiences online"; C&I = "Confidence and identification"; I&P = "Information and presentation"; U&M = "Understanding and motivation"; CSHEO = "Comfort with sharing health experiences online"; USHEO = "Usefulness of sharing health experiences online"; M&CA = "Motivation and confidence to act"; ID = "Identification"; Grade = Overall satisfaction; NPS = Net Promoter Score; EQ-5D = EuroQol-5D; FCCHL = Functional, Communicative and Critical Health Literacy scale

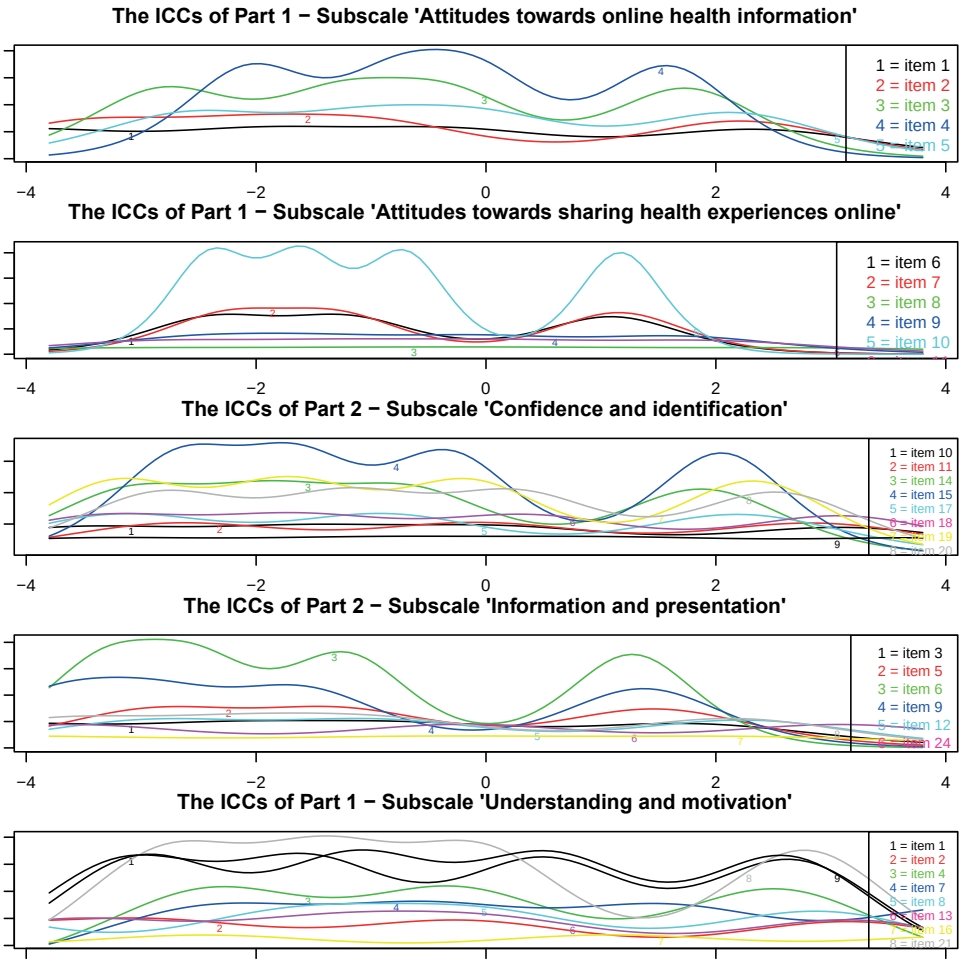

Figure 5.1. Study Sample 1: Item Information Curves of original subscales.

Figure 5.2 shows the item information curves of the modified subscales. Of part 1, the information of the subscale *comfort with sharing health experiences online* was rather low across the entire latent trait spectrum. For the subscale "Usefulness of sharing health experiences online" information was high on certain points of the latent trait spectrum, but all three items overlap almost completely. Of part 2, the subscale *motivation and confidence to act* showed a good range of information across latent trait levels. However, 3 items hardly contributed information (items 1, 7, and 13). The subscale *information and presentation* still suffered from multiple items adding little information, as well as a lot of overlap. Lastly, the subscale *identification* showed a good range of information as well as high peaks for all 3 items, but still a lot of overlap between items on information range.

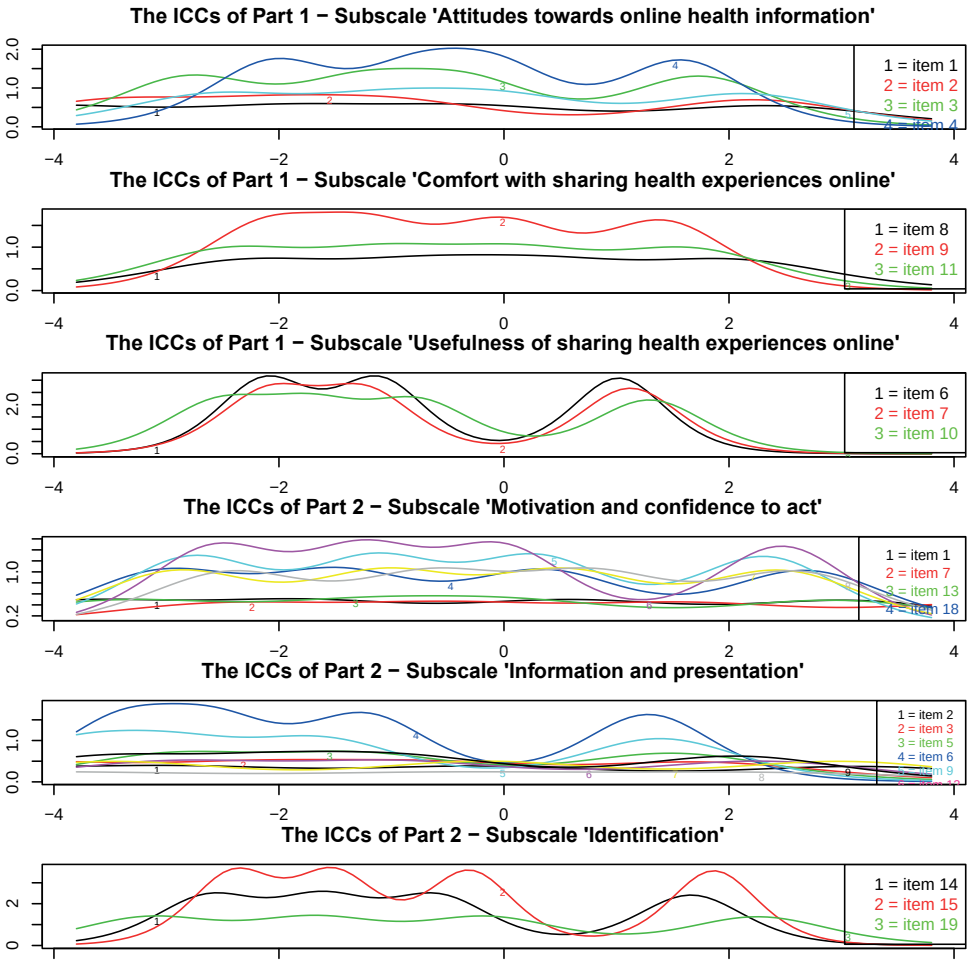

Figure 5.2. Study Sample 1: Item Information Curves of modified subscales.

5.2.3 Study Sample 2

5.2.3.1 Structural Validity

A CFA was run with the 2 original subscales as first-order factors. This model had a bad fit (CMIN = 8.13, AGFI = .829, CFI = .893, TLI = .863, SRMR = .054, and RMSEA = .113 [.102 - .124]). A second CFA was run with the modified factor structure found in study 1. This model had a barely acceptable fit (CMIN = 7.37, AGFI = .849, CFI = .909, TLI = .878, SRMR = .049, and RMSEA = .107 [.096 - .118]). An EFA using Oblimin rotation was run with 3 factors, to determine possible deviations from the 3 subscales found study 1. This model had a good fit (CMIN = 4.86, AGFI = .978, CFI = .966, TLI = .926, SRMR = .025, and RMSEA = .084 [.069 - .098]). The 3 factors

(*Table 5.2C*) were identical to the subscales found in study 1, except for item 8 loading on both subscales concerning the sharing of health experiences online.

5.2.3.2 Internal Consistency

Table 5.3C shows the internal consistency of the original factor structure and the modified factor structure. All values were acceptable (omega > .70), and comparable between both factor structures.

5.2.3.3 Divergent Validity

For both the original and the modified factor structure, only the subscale *Attitudes toward online health information* showed a significant correlation with the FCCHL (*Table 5.6C*). However, this correlation is small enough to be acceptable for divergent validity (r < .15).

5.2.3.4 Graded Response Model

Figure 5.3 shows the item information curves for the original subscales of part 1. A number of items do not provide much extra information over the others: items 5, 8, and 9. Figure 5.4 shows the item information curves of the modified subscales of part 1. The information of the subscale *comfort with sharing health experiences online* showed large dips on certain levels of ability. For the subscale *usefulness of sharing health experiences online* information was high on certain points of the latent trait spectrum, but the items overlap a great deal.

5.2.4 Study Sample 3

5.2.4.1 Structural Validity

A CFA was run with the 3 original subscales as first-order factors. This model had a slightly below acceptable fit (CMIN = 5.568, AGFI = .717, CFI = .811, TLI = .792, SRMR = .092, and RMSEA = .093 [.089 - .098]). A second CFA was run with the 3 modified subscales found in study sample 1. This model had an acceptable fit (CMIN = 4.447, AGFI = .828, CFI = .889, TLI = .873, SRMR = .075, and RMSEA = .081 [.075 - .087]). An EFA using Oblimin rotation was run with 3 factors and including the items that were deemed problematic in Study sample 1, to determine whether including them would result in a better fit. This model had a good fit (CMIN = 2.496, AGFI = .990, CFI = .948, TLI = .932, SRMR = .029, and RMSEA = .053 [.048 - .059]).

In the EFA, Items 8, 11, and 17 showed no problematic cross loadings. Items 4, 10, and 16 did show problematic cross loadings, but not as extreme as in study sample 1 (*Table 5.2D*). Items 8, and 10 were found to load most highly on the factor

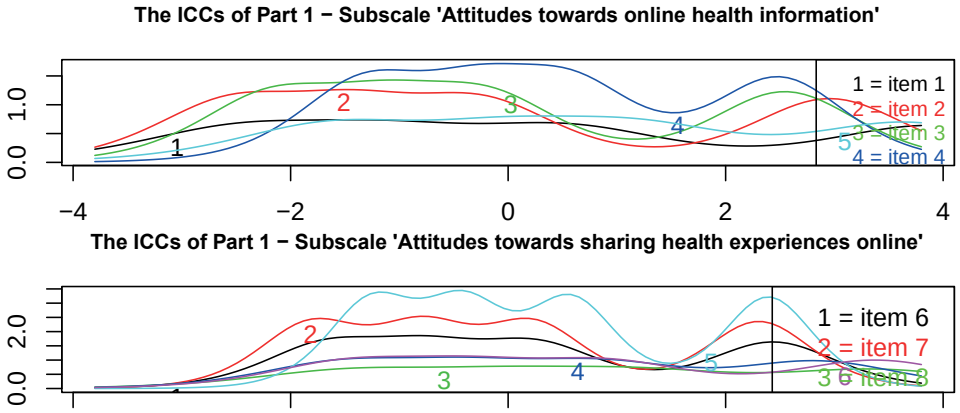

Figure 5.3. Study Sample 1: Item Information Curves of original subscales.

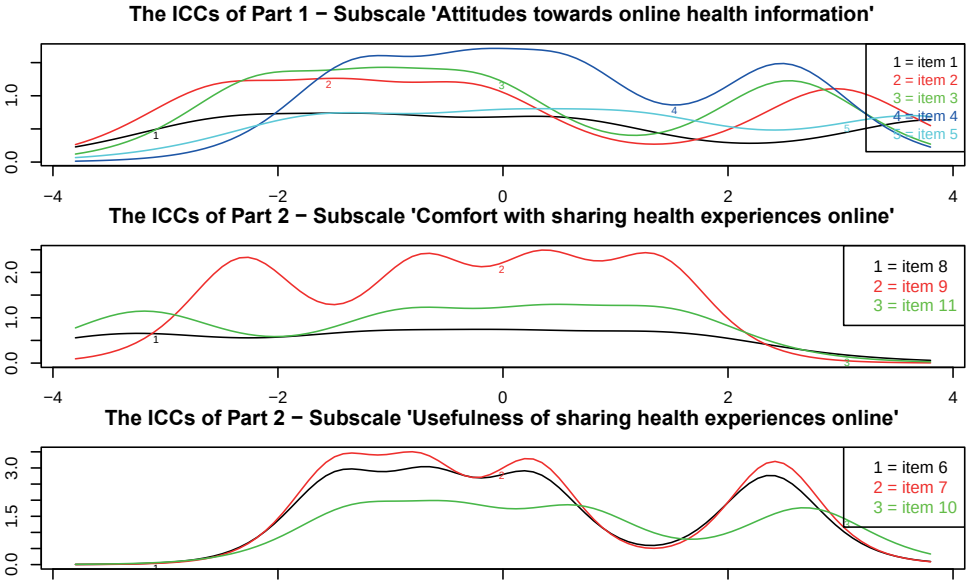

Figure 5.4. Study Sample 1: Item Information Curves of modified subscales.

representing *motivation and confidence to act*. Items 4, 11, 16, and 17 were found to load most highly on the factor representing *information and presentation*. Beyond the problematic items, only one item loaded differently than in study sample 1: Item 20 loaded as highly on the factor representing *motivation and confidence to act* (on which it loaded in study sample 1) as it did on the factor representing *identification*.

5.2.4.2 Internal Consistency

Table 5.3D shows the internal consistency of the original factor structure, the modified factor structure without previously problematic items, and the modified factor structure with previously problematic items, respectively. The internal consistency of the modified factor structure with previously problematic items is represented by Cronbach's alpha instead of omega, as omega is based on factor variance and unsuitable for factor structures fit based on EFAs. All values, except for the original subscale *information and presentation* (omega = .65), were acceptable, and comparable between the 3 factor structures.

5.2.4.3 Convergent Validity

Both the original and modified subscales correlated significantly with the System Usability Scale, Net Promoter Score and grade questions (*Table 5.6D*). All correlations were acceptable for convergent validity ($r > .30$), except for the original subscale *confidence and identification* with the SUS ($r = .29$), and the modified subscale *identification* with the SUS ($r = .12$).

5.2.4.4 Graded Response Model

Figure 5.5 shows the item information curves for the original subscales of part 2. With a large number of items per scale, there was a good range of information across latent trait levels. Some items did not add much to the information provided by other items: items 3, 8, 10, 11, 17, 21, 24, 25, and 26. Notably, items 8, 10, 11, and 17 were items that were judged problematic in study 1. Figure 5.6 shows the item information curves of the modified subscales of part 2. The subscale *motivation and confidence to act* showed a good range of information across latent trait levels. However, 3 items hardly contributed information: items 1, 7, and 13. The subscale *information and presentation* still suffered from multiple items adding little information, as well as a lot of overlap. Finally, the subscale *identification* showed a good range of information as well as high peaks for all 3 items, but still a lot of overlap.

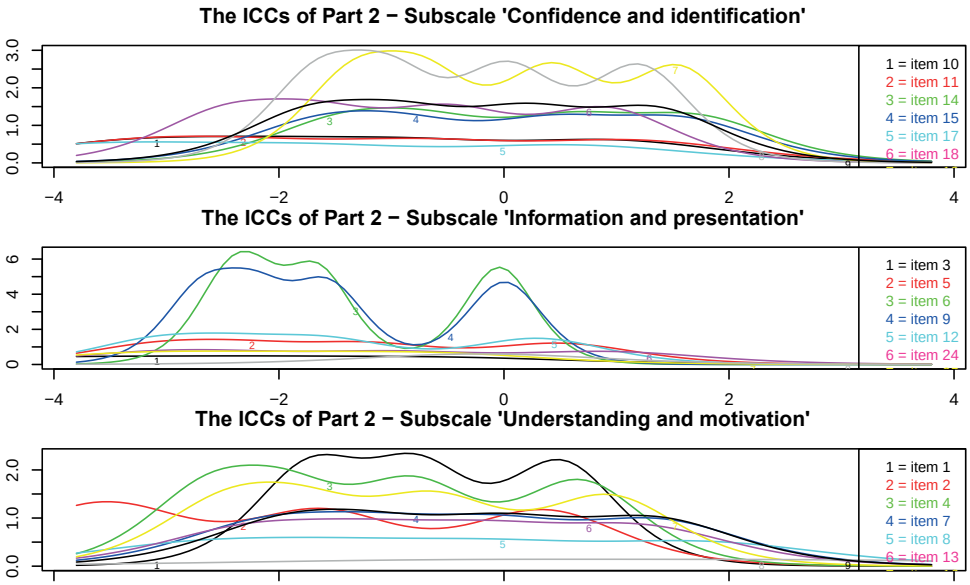

Figure 5.5. Study Sample 1: Item Information Curves of original subscales.

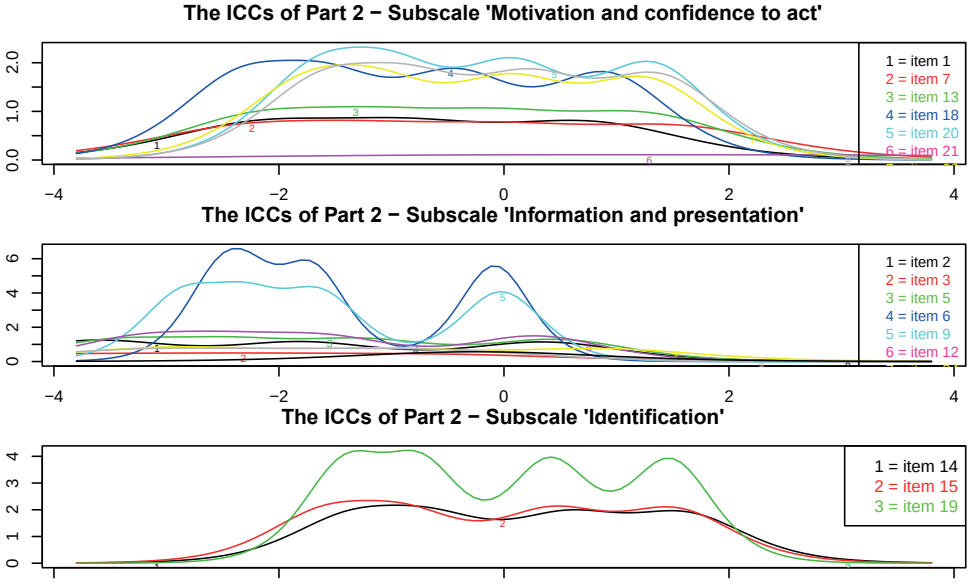

Figure 5.6. Study Sample 1: Item Information Curves of modified subscales.

5.3 Discussion

5.3.1 Principal Findings

In this study the eHIQ was translated into Dutch and the measurement properties were investigated. Feasibility was good: more than 94% of participants in the main study completed the eHIQ-NL. The eHIQ-NL showed a different factor structure compared with the original English version. Part 1 of the eHIQ-NL consists of 3 subscales: *attitudes towards online health information* (5 items), *comfort with sharing health experiences online* (3 items), and *usefulness of sharing health experiences online* (3 items). Part 2 of the eHIQ-NL consists of 3 subscales: *motivation and confidence to act* (10 items), *information and presentation* (13 items), and *identification* (3 items). These factor structures were replicated in subsequent samples, and altogether showed acceptable to good internal consistency, test-retest reliability, and construct validity.

5.3.2 Limitations

Limitations of this study are some underperforming measurement properties of the modified factor structure. In particular test-retest reliability for *comfort with sharing health experiences online and usefulness of sharing health experiences online* (ICC = .62, and .53, respectively) was below acceptable threshold. Notably, the original subscale comprised of these 2 subscales *attitudes towards sharing health experiences online* also underperformed on test-retest reliability (ICC = .63).

Furthermore, the correlations testing convergent validity were small in the main study ($r < .30$), as well as some smaller correlations in study sample 3 for the subscales *confidence and identification* ($r = .29$), and the modified subscale *identification* ($r = .12$). We recognize that this may be because of subpar a priori hypotheses in regard to the EQ-5D (study 1) and the SUS (study 3). The reasoning for these hypotheses was somewhat tenuous. For the first sample we expected the specific eHealth application Kanker.nl to provide useful information for patients with issues regarding daily activities and anxiety/depression resulting in a correlation between a higher score on these issues and eHIQ scores. For the third sample, we expected a higher usability score to be correlated to higher eHIQ scores, but we recognize that the subscales *confidence and identification* and *identification* may be theoretically unrelated to usability. Further research is necessary to further investigate test-retest reliability and construct validity of the eHIQ-NL. Future validations in different nationalities and different patient populations may shed more light on these measurement properties.

5.3.3 Comparison With Prior Work

The findings of this study do not entirely match the findings of the original validation of the eHIQ for the British population [36]. The differences may be the results of a number

of differences between the current and previous validation studies. The first explanation is that in the translation of the questionnaire the meaning of some items may have changed. Although we followed a strict protocol for the translation, this explanation cannot be ruled out.

The second explanation can be found in the use of a different study populations. The original validation study presented the eHIQ to a range of health groups, who were not necessarily eHealth users at the time of the study. The participants in the original validation study were invited to the laboratory, and were briefly (at least 15 min) acquainted with an eHealth application relevant to their personal health situation [36]. This study presented the eHIQ-NL to eHealth users who were familiar with the application under investigation (study samples 1 and 3) and non-current eHealth users. Furthermore, the current validation study presented the eHIQ-NL only to cancer patients (study samples 1 and 2) and patients with musculoskeletal disorders (study sample 3). As such, the populations differ quite a bit beyond nationality.

The results of this study presents complexities to which subscales should be adhered to. Depending on the goal of the user of the eHIQ-NL, we propose either the use of the original subscales or the use of subscales based on the factor structure which were found in the Dutch population. If the user of the eHIQ-NL wants to be able to compare their results to international samples, the original subscales should be adhered to. The caveat is that one cannot be sure of the structural validity using this method, and we recommend factor analysis to back up any interpretation. If the user does not intend to compare their results to international samples, the use of subscales derived from our results is recommended. The caveat is that this makes the result incomparable to international samples, and as such the data could not be used in future re-analysis of an international nature.

5.3.4 Conclusions

Nevertheless the limitations specified above, the eHIQ-NL shows a consistent factor structure, sufficient internal consistency, and mostly sufficient test-retest reliability and construct validity. The eHIQ-NL is a valid and reliable tool for measuring attitudes of eHealth users, and can be implemented using the original subscales or modified subscales depending on the nature of the research question. Interested users can contact Oxford Innovations (healthoutcomes@innovation.ox.ac.uk) for a license to use the eHIQ.


6

Chapter 6

Symptom Cluster in Cancer Survivors

This chapter was submitted as Neijenhuijs, K. I., Peeters, C. F. W., van Weert, H., Cuijpers, P., & Verdonck-de Leeuw, I. M. (2019). Symptom clusters among cancer survivors: what can machine learning techniques tell us?

Abstract

Purpose: Knowledge regarding symptom clusters may inform targeted interventions. The current study investigated symptoms clusters among cancer survivors, using machine learning techniques on a large data set.

Patients and methods: Data were used of cancer survivors who used a fully automated online application 'Oncokompas' that supports them in their self-management by 1) monitoring their symptoms through patient reported outcome measures (PROMs); and 2) providing tailored feedback on their scores with a personalized overview of supportive care options, aiming to reduce symptoms burden and improve health-related quality of life. In the present study, data on 26 generic symptoms (physical and psychosocial) were used. Results of the PROM of each symptom are presented to the user as a no well-being risk, moderate well-being risk, or high well-being risk score. Data of 1032 cancer survivors were analysed using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) on high risk scores and moderate-to-high risk scores separately.

Results: When analysing the high risk scores, seven clusters were extracted: one main cluster which contained most frequently occurring physical and psychosocial symptoms, and six subclusters with different combinations of these symptoms. When analysing moderate-to-high risk scores, three clusters were extracted: two main clusters were identified, which separated physical symptoms (and their consequences) and psycho-social symptoms, and one subcluster with only body weight issues.

Conclusion: There appears to be an inherent difference on the co-occurrence of symptoms dependent on symptom severity. Among survivors with high risk scores, the data showed a clustering of more connections between physical and psycho-social symptoms in separate subclusters. Among survivors with moderate-to-high risk scores, we observed less connections in the clustering between physical and psycho-social symptoms.

Cancer survivors experience a myriad of symptoms rooted in physiology caused by the disease itself or caused by treatment [12]. Problems in the psychosocial domain are also prevalent [10,13,14]. Many of these symptoms and problems may co-occur and are likely interrelated. For example, sleep problems have been identified as both a risk factor and as a symptom of depression in both cancer [230] and non-cancer populations [231]. In cancer patients, subjective cognitive functioning has been associated with depression, anxiety, and fatigue [232]. Furthermore, Problems with sexual health have been related to body image issues, depression, and anxiety [233]. Fatigue has been associated with pain, sleep issues, and depression; and nausea with vomiting [46].

Such interrelated symptoms are referred to as symptom clusters, and knowledge regarding such symptom clusters may inform targeted interventions [48]. Some studies have set out to empirically determine symptom clusters using various types of cluster analyses. In 2011, a systematic review identified 47 studies that statistically investigated cluster symptoms in cancer patients [46]. A number of clusters repeatedly showed up: (i) a fatigue-depression-pain cluster, (ii) a nausea-vomiting cluster, (iii) a depression-anxiety-insomnia cluster. However, the authors noted that these (and other) clusters seem heavily influenced by the population studied (tumour type, stage, treatment modality and intent), symptom assessment method, and statistical method used. Also, many of these studies were limited in scope in terms of sample size, number of symptoms investigated, or the type of analysis that was used.

Another systematic review was performed for studies up to 2016 which focused on cancer patients receiving primary or adjuvant chemotherapy, specifically [47]. Nineteen studies were included, and a few consistently appearing clusters were identified: (i) a nausea-vomiting cluster, (ii) a psychological symptom cluster, and (iii) a "sickness behaviour" (pain-fatigue-insomnia-lack of appetite) cluster. Noteworthy is that the individual symptoms in each of these clusters were not necessarily consistent between studies.

In 2015 an international expert panel regarding "Advancing Symptom Science Through Symptom Cluster Research" was formed [48]. This panel was subdivided into five groups: (i) defining characteristics of symptom clusters, (ii) identification of priority symptom clusters and underlying mechanisms, (iii) measurement of symptom clusters, (iv) targeted interventions for symptom clusters, and (v) new analytic strategies for symptom cluster research. In line with the aforementioned review [46], the first expert group concluded that there is little consistency in the number and types of symptom clusters identified in cancer patients/survivors. This expert group defined a number of directions for future research in defining symptom clusters. In particular, they stated a

need for "the establishment of a common conceptual framework and approach for the evaluation of measurement of symptom clusters" and "the evaluation of the potential to use large data sets and electronic health records to evaluate symptom clusters". The fifth expert group also defined some direction for future research, one of which was "apply new analytic techniques to symptom cluster research". The current study hopes to contribute to all three of these directions.

The aim of this study is to investigate symptom clusters among cancer survivors, by analysing a large dataset of self-reported symptoms using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [234], a recent development in cluster analysis. The results of the present study will contribute to the establishment of a conceptual framework and approach for the evaluation of symptom cluster measurements.

6.1 Methods

6.1.1 Study population

The sample consisted of users of the eHealth application Oncokompas [8]. In total, data of 1032 cancer survivors were used, who consented that their data were used for research purposes. 715 users of Oncokompas who were referred through a healthcare provider in routine care, 191 cancer survivors who were invited to participate in a randomized controlled trial (RCT) investigating the efficacy of Oncokompas [8], 72 colon cancer survivors who were invited to participate in a multi-centre RCT [235], and 54 breast cancer survivors who were invited to participate in a pilot on the feasibility of Oncokompas [58].

6.1.2 Materials

Oncokompas is a fully automated online application that supports cancer survivors in their self-management by 1) monitoring their health-related quality of life (HRQOL) and (cancer-generic and tumour-specific) symptoms; and 2) obtaining tailored feedback on their scores with a personalized overview of supportive care options, with the aim to reduce symptoms burden and improve HRQOL [8]. Oncokompas covers in total 46 topics on five generic domains applicable for all cancer survivors: physical, psychological, and social HRQOL, healthy lifestyle, and existential topics; and 29 tumour-specific topics for survivors of breast cancer, colorectal cancer, head and neck cancer, and lymphoma. Users can choose between topics. In the current study, only the generic topics were used. Oncokompas consists of three components: 'Measure', 'Learn', and 'Act'. For the current study, only the Measure component is of interest. The Learn and Act component are detailed elsewhere [8]. In the Measure component, users can independently complete patient reported outcome measures (PROMs) targeting

the selected topic(s). On each of the selected topics, the user receives a green (no well-being risk), orange (moderate well-being risk), or red (high well-being risk) outcome. The current study focuses on 26 of the 46 generic topics, as these 26 topics represent physical or psycho-social symptoms that often occur based on literature [19]. Table 6.1 details the symptoms and PROMs that were used in the analysis, as well as the possible colour outcomes on each PROM. Each symptom consists of one or multiple PROMs, which were selected by the project team in collaboration with expert teams and based on Dutch practical guidelines (from the Netherlands Comprehensive Cancer Organisation) and literature searches [9].

Table 6.1. Overview of Oncokompas topics.

Topics	PROM	Possible scores
Contact with doctor	EORTC IN-PATSAT32	Green; Orange
Dedication to work	Visual Analogue Scale	Green; Orange
Smoking	Oncokompas expert-based questionnaire	Green; Orange; Red
Alcohol use	Alcohol 5-shot	Green; Orange; Red
Relaxation	Perceived Stress Scale	Green; Orange; Red
Physical activity	Oncokompas expert-based questionnaire	Green; Orange
Body weight	BMI & Short Nutritional Assessment Questionnaire	Green; Orange; Red
Physical limitations daily life	Patient Specifieke Klachtenlijst (Dutch-specific)	Green; Orange; Red
Insomnia	Insomnia Severity Index	Green; Orange; Red
Fatigue	Numeric Rating Scale	Green; Orange; Red
Pain	Numeric Rating Scale	Green; Orange; Red
Constipation	Numeric Rating Scale	Green; Orange; Red
Diarrhea	Numeric Rating Scale	Green; Orange; Red
Lack of appetite	Numeric Rating Scale	Green; Orange; Red
Nausea or vomiting	Numeric Rating Scale	Green; Orange; Red
Shortness of breath	Numeric Rating Scale	Green; Orange; Red
Hearing problems	Caron hearing questionnaire	Green; Orange; Red
Tinnitus	Oncokompas expert-based questionnaire	Green; Orange; Red
Psychological complaints	Hospital Anxiety and Depression Scale	Green; Orange; Red
Memory / concentration	SF-36 'cognitive functioning'	Green; Orange; Red
Social life	De Jong-Gierveld Loneliness Scale	Green; Orange; Red
Financial problems	EORTC QLQ-C30 'financial problems'	Green; Orange; Red
Intimacy and sexuality	Female Sexual Function Index (women) / International Index of Erectile Function (men)	Green; Orange; Red
Body image	Body Image Scale	Green; Orange; Red
Relationship with partner	Dyadic Adjustment Scale Short Form	Green; Orange; Red
Relationship with children	Vragenlijst Gezinskernmerken Short Form (Dutch-specific)	Green; Orange; Red

6.1.3 Data analysis

Data of cancer survivors who used Oncokompas up to April 29th 2019 was used. Users can fill in Oncokompas more than once. To remove within-user variance, when users had filled in Oncokompas more than once, one random time point was selected of the user. All analyses were run in R version 3.5.3 [222], or in Python version 3.7.1 [236]. Two types of analyses were used: network analysis and cluster analysis.

For the cluster analyses, HDBSCAN [234] was performed using the hdbscan library in Python [237]. HDBSCAN separates a dataset into clusters of high and low density. HDBSCAN is an extension of the DBSCAN clustering algorithm [238], where HDBSCAN is capable of identifying clusters of varying densities and is more robust to parameter selection [237]. This makes the HDBSCAN algorithm particularly useful in separating smaller subclusters (with higher densities) from larger clusters (with lower densities). Data points that do not fit into any of the identified clusters are labelled as noise by the algorithm. The Jaccard distance metric was used due to the categorical nature of our measurement of a symptom. The minimum points required to form a cluster (minimum cluster size) was set to 26 (number of modules). Because we were interested in subclusters of symptoms, the minimum sample was set to 1, and leaf clustering was used for cluster selection. These parameters prioritize the extraction of multiple smaller rather than larger clusters.

The network analyses were performed using the tidygraph [239] and ggraph [240] packages. The network graphs were nondirectional, and edges were calculated as the raw number of connections between nodes (i.e. the occurrence of symptom-pairs among the same patient). Weighted degree centrality was calculated using the edges as weights.

One analysis was run on only high risk (red) scores as the definition of a symptom being present, as these scores are based on cut-off scores with most empirical evidence. Three symptoms on which a user cannot score red (see (*Table 6.1*)) were excluded for this analysis. A second analysis was run on moderate-to-high risk (orange and red) scores as the definition of a symptom being present, which also included the previously excluded symptoms.

6.2 Results

6.2.1 Patient characteristics

Table 6.2 shows the patient characteristics. The mean age was 61.5 years (range 25 - 88), the majority was female (68%), approximately half was treated for breast cancer (49%), most had completed treatment (60%), and most patients were treated with surgery (79%).

Table 6.2. Descriptive statistics.

	Mean	SD		N	%
Age	61	11			
Gender			Female	701	67.99%
			Male	330	32.01%
Education			Elementary school	24	2.33%
			High school	164	15.91%
			Vocational education	590	57.23%
			College	115	11.15%
			University	107	10.38%
			Post-doctoral	23	2.23%
			Other	8	0.78%
Cancer type			Breast cancer	504	48.88%
			Colon cancer	182	17.65%
			Lymphoma	73	7.08%
			Head and neck cancer	60	5.82%
			Rectal cancer	40	3.88%
			Other	39	3.78%
			Lung cancer	31	3.01%
			Prostate cancer	29	2.81%
			Gynecologic cancer	20	1.94%
			Bladder or kidney cancer	17	1.65%
			Skin cancer	11	1.07%
			Blood cancer	9	0.87%
			Esophageal cancer	5	0.48%
			Brain cancer	4	0.39%
			Pancreatic or liver cancer	4	0.39%
			Stomach cancer	3	0.29%
Treatment status			Treatment completed	614	59.55%
			Currently being treated	172	16.68%
			Not yet treated	94	9.12%
			Unknown	80	7.76%
			No treatment	71	6.89%
Treatment type			Surgical	743	79.04%
			Chemotherapy	103	10.96%
			Radiation	38	4.04%
			Chemoradiation	24	2.55%
			Other	10	1.06%
			Hormone therapy	7	0.74%
			Wait-and-see	7	0.74%
			Immunotherapy	4	0.43%
			Unknown	4	0.43%

6.2.2 High risk score analysis

In the high risk score analysis, seven clusters were extracted. A total of 393 data points were deemed noise, which amounted to 19.31% of the data, which is a non-negligible amount.

The cluster profiles are presented in Table 6.3. The cell numbers represent how many patients with a certain symptom were present in any given cluster. The largest cluster (cluster 7), represents a "general sickness cluster", encompassing patients who suffer from most symptoms that were represented in the data set. Next are two clusters that represent patients who experience one symptom almost exclusively: a psychological complaint cluster (cluster 1), and a physical limitations cluster (cluster 2). Cluster 3 represents patients who mainly experience symptoms on body weight, alcohol use, and social life, while cluster 4 represents patients who mainly experience symptoms on physical limitations, intimacy/ sexuality, and body weight. cluster 5 is the second-largest cluster in regards to number of symptoms, and represents patients who experience psychological symptoms and various physical symptoms. Lastly, cluster 6 represents patients who experience psychological complaints, problems with relaxation, and social life.

Table 6.3. High risk cluster profiles.

Symptom	Noise	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Psychological complaints	225	65		1	6	36	30	52
Fatigue	114			1	6	21	1	48
Physical limitations daily life	100		27		18	33	5	46
Memory / concentration	53					1	3	26
Relaxation	123					1	31	25
Social life	82			12		9	29	24
Insomnia	66			1	1	3	2	21
Pain	64			1		10	1	19
Intimacy and sexuality	74			1	15	11		15
Hearing problems	107				4	26	1	15
Shortness of breath	47				2	15	1	9
Body weight	35			22	14	3		8
Constipation	25						3	3
Smoking	41			3		6	1	2
Diarrhea	33							2
Body image	15	3						1
Relationship with partner	21					1	4	1
Alcohol use	22			12	2		2	
Lack of appetite	23							
Relationship with children	19							
Nausea or vomiting	18							
Financial matters	17							
Tinnitus	6							

Note: The cell numbers represent how many patients with a certain symptom were present in any given cluster. Noise indicates data points that do not fit into any of the identified clusters).

Figure 6.1 shows the network plot of the main analysis. The plot shows both the main cluster of each symptom (the cluster in which the symptom is most frequent), as well as the subcluster of each symptom (the cluster in which the symptom is second-most frequent, with a minimum frequency of 5). Psychological complaints and physical limitations have the highest weighted degree centrality and are connected to nearly all symptoms. The intra-cluster connections range from large (mostly connections originating from psychological complaints or physical limitations), to moderate (most intra-cluster connections), to small (mostly connections originating from fringe symptoms with less neighbours).

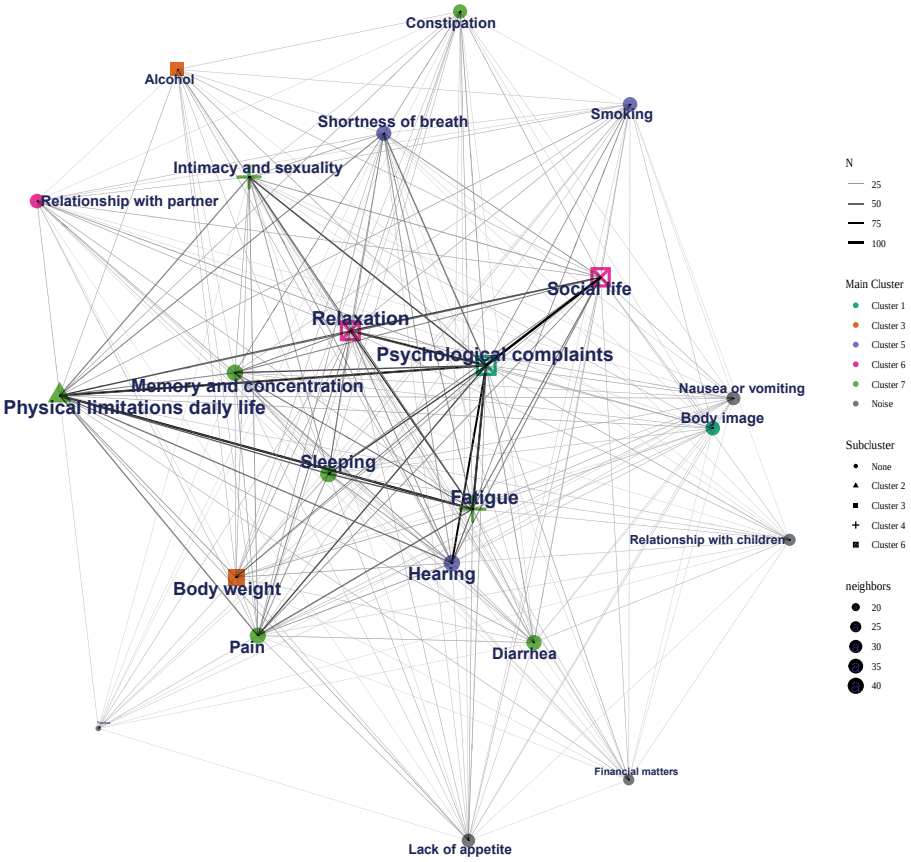

Figure 6.1. Network plot high risk score analysis.

6.2.3 Moderate-to-high risk score analysis

In the moderate-to-high risk score analysis, three clusters were extracted. A total of 579 data points were deemed as noise, which amounted to 39.25% of the data, which is a high amount.

The cluster profiles are presented in Table 6.4. One small cluster was extracted with patients who only experienced symptoms on body weight (Cluster 1). Two large clusters emerged: A lifestyle and psychosocial cluster (cluster 2), and a physical symptoms cluster (Cluster 3).

Table 6.4. Moderate-to-high risk cluster profiles.

Symptom	Noise	Cluster 1	Cluster 2	Cluster 3
Physical activity	257		175	130
Relaxation	105		88	67
Insomnia	99		5	67
Contact with doctor	75		46	66
Shortness of breath	79		1	65
Physical limitations daily life	100		2	51
Fatigue	136		3	50
Social life	58		57	43
Alcohol use	147		98	39
Tinnitus	85		3	39
Pain	59			38
Financial matters	71		52	36
Lack of appetite	26			25
Intimacy and sexuality	58		2	24
Relationship with partner	39		19	18
Psychological complaints	96		74	16
Constipation	26		1	16
Nausea or vomiting	16			16
Hearing problems	26		3	11
Memory / concentration	13			11
Body weight	62	18	123	10
Relationship with children	36		32	10
Body image	16			10
Dedication to work	28		1	8
Diarrhea	5			4
Smoking	15		7	2

Note: The cell numbers represent how many patients with a certain symptom were present in any given cluster. Noise indicates data points that do not fit into any of the identified clusters).

Figure 6.2 shows the network plot of the sensitivity analysis. The plot does not show the body weight cluster, as this symptom was strongly incorporated into the psychosocial and lifestyle cluster. Psychological complaints, physical limitations, physical activity,

relaxation, and fatigue have the highest weighted degree centrality and are connected to nearly all symptoms. The intra-cluster connections range from large (mostly connections originating from the symptoms with high weighted degree centrality), to moderate (most intra-cluster connections), to small (mostly connections originating from fringe symptoms with less neighbours). This network analysis shows more inter-cluster connections than the main analysis. Strong connections exist between psychological complaints (cluster 2) and fatigue (cluster 3), and physical limitations (cluster 3). Moderate connections exist between psychological complaints (cluster 2) and intimacy/sexuality (cluster 3), pain (cluster 3), and insomnia (cluster 3). Moderate connections exist between physical limitations (cluster 3) and social life (cluster 2), physical activity (cluster 2), relaxation (cluster 2), and social life (cluster 2).

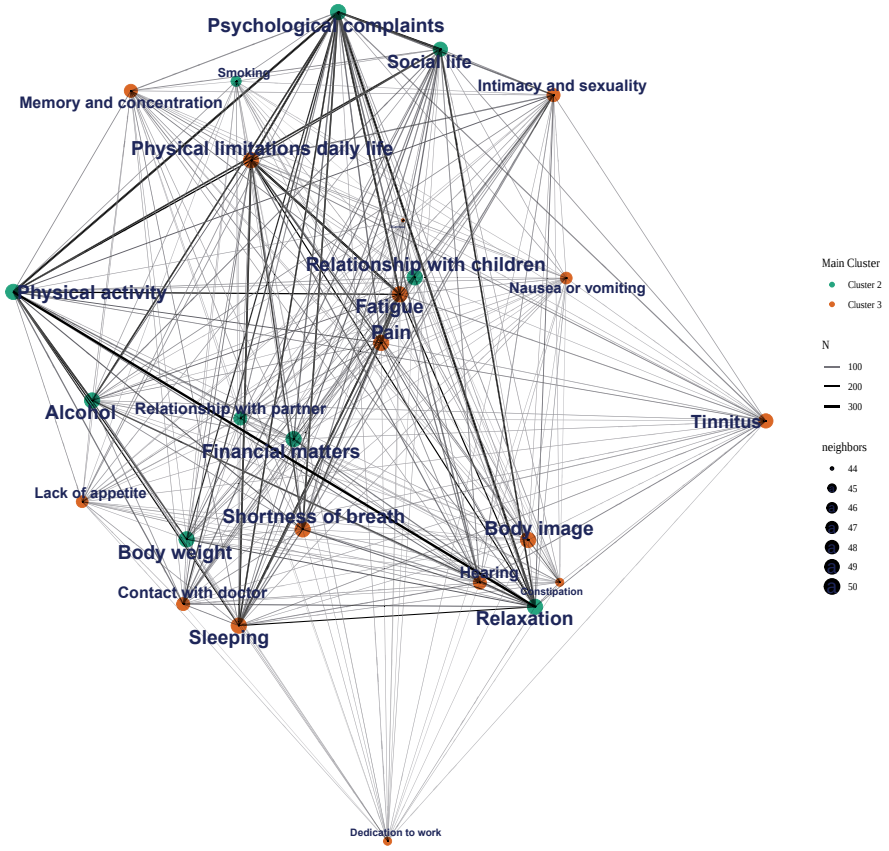

Figure 6.2. Network plot moderate-to-high risk score analysis.

6.3 Discussion

In this explorative study we used HDBSCAN to extract symptom clusters based on scores on PROMs of cancer survivor users of the eHealth application Oncokompas. Different clusters appeared when analysing the high risk scores of patients, versus when analysing moderate-to-high risk scores. When analysing patients showing high-risk scores, we found one overarching cluster containing most symptoms measured, as well as six subclusters. When analysing both patients showing moderate-to-high risk scores, we found two overarching clusters: one representing psychosocial symptoms, and one representing physical symptoms and their health consequences.

This study was explorative in nature, and while the symptoms used in the current study do not entirely line up with all symptoms reported on by previous research [46,47,230,232,233], it is of interest to see whether symptom clusters identified in our dataset line up with the symptom clusters previously identified.

First, a fatigue-depression-pain cluster [46,47] has been reported. We found fatigue, psychological complaints, and pain to be clustered together in the "general sickness" cluster as well as in the "physical symptoms and consequences" cluster in the high risk score analysis. And while fatigue and pain clustered in the "physical symptoms and consequences" cluster in the moderate-to-high risk score analysis, they were not clustered with psychological complaints.

Second, previous literature showed evidence for a depression-anxiety-insomnia cluster [46,230]. Insomnia and psychological complaints were clustered together in the "general sickness" cluster, and showed a moderate connection in the network analysis (weight = 76) in the high risk score analysis. In the moderate-to-high risk score analysis insomnia and psychological complaints were not clustered together, but did show a a strong connection in the network analysis (weight = 219).

Third, a psychological symptom cluster was found in multiple previous studies [46,47]. In the high risk score analysis we found both a "psychological complaints" and "psychosocial" cluster. However, in the moderate-to-high risk score analysis we found a broader psychosocial and lifestyle cluster.

Fourth, a pain-fatigue-insomnia-lack of appetite clusters was reported [47]. In the high risk score analysis, pain, fatigue, and sleeping issues clustered together in the "general sickness" cluster; and pain and fatigue clustered in the "physical symptoms and consequences" cluster. Meanwhile, in the moderate-to-high risk score analysis, pain, fatigue, sleeping issues, and lack of appetite clustered in the "physical symptoms and consequences" cluster.

Fifth, a specific association between cognitive functioning and psychological distress was found in previous literature [232]. In the high risk score analysis, memory/concentration and psychological complaints were clustered together in the "general sickness" cluster, and show a moderate connection in the network analysis (weight = 66); while in the moderate-to-high risk score analysis memory/concentration and psychological complaints were not clustered together, but did show a moderate connection in the network analysis (weight = 98).

Sixth, literature showed a specific association between memory/concentration and fatigue [232]. In the high risk score analysis, memory/concentration and fatigue clustered in the "general sickness" cluster, and showed a small to moderate connection in the network analysis (weight = 54). In the moderate-to-high risk score analysis, memory/concentration and fatigue clustered in the "physical symptoms and consequences" cluster, and showed a moderate connection in the network analysis (weight = 93).

Seventh, a specific association between sexual problems and body image has been reported [233]. Intimacy/sexuality and body image were not clustered together in the high risk score analysis, as body image was not part of any cluster, and showed a very weak connection in the network analysis (weight = 5). In the moderate-to-high risk score analysis, intimacy/sexuality and body image clustered in the "physical symptoms and consequences" cluster, but showed a weak connection in the network analysis (weight = 22). These results may be explained by the fact that we have very few patients in the main data set with body image problems.

Eight and last, a specific association between sexual problems and psychological distress was found previously [233]. In the high risk score analysis, intimacy/sexuality and psychological complaints were clustered together in the "general sickness" cluster as well as in the "physical symptoms and consequences" cluster, and showed a moderate connection in the network analysis (weight = 73). In the moderate-to-high risk score analysis, intimacy/sexuality and psychological complaints were not clustered together, but did show a strong connection in the network analysis (weight = 159).

These results show that many of the previously reported (sub)clusters were found in the high risk score analysis, but not in the moderate-to-high risk score analysis. There may be an inherent difference on the co-occurrence of symptoms dependent on symptom severity. For patients with higher symptom severity, we observed more connections between the physical and psycho-social symptoms; while for patients with lower symptom severity, we observed less connections between physical and psycho-social symptoms. This motivates the question of causality: Do patients with higher

severity of physical symptoms develop higher severity of psycho-social symptoms, vice versa, or are both higher severities developed in tandem due to a third causal force? The statistical methods we used are associative, intending to identify clusters of co-occurring symptoms which do not necessarily share the same aetiology [48]. As such, we cannot offer an answer to the question of causality. But future research could use methodology more suited for such investigations.

It has been suggested that cluster symptoms are not the same across differing cancer diagnoses [47]. The current study did not perform subgroup analyses between cancer diagnoses, as the corresponding sample sizes would not have been sufficient for all diagnoses. In future research, after Oncokompas has attracted more users of differing diagnoses, such subgroup cluster analyses may provide further insights into this possibility.

One strength, but simultaneously a limitation, of this study is the use of different measurement tools for each separate symptom. It has been argued that a standardization of how to measure symptoms for use in classifying symptom clusters is necessary for reproducible and valid interpretations [48]. However, the use of multiple (standardized and validated) measurement instruments creates the possibility to analyse many more symptoms than would be possible when using only one standardized measurement tool.

There are two further limitations in regards to the way data was analysed. First by focusing on the extraction of smaller clusters we found a high noise count in the cluster analyses. This indicates that there may be other (likely larger) clusters that could be extracted with other parameter settings. The amount of noise data points could also likely be reduced by using an algorithm that searches for the optimal minimum cluster size. While such settings were not judged optimal to answer our particular research question, such analyses could improve the fit of the model to the data. Second, for users that filled in Oncokompas more than once, we chose a random data row to ensure that we did not increase bias in our data set. Another dataset (e.g. selecting the data row which includes most moderate-to-high risk scores) could have produced a differently informed result.

It is of interest to see whether the cluster symptoms found in the current study can be replicated in samples using other measurement techniques, as well as other analysis techniques. A form of standardization on both has been argued [48], but which measurement and analysis techniques are most appropriate and should be the standard has not yet been firmly concluded. Through replication, information on comparability can be gained. Future studies could also expand on the current study through the use

of causal modelling techniques to investigate possible aetiological connections between symptoms, as well as by using subgroup analyses for differing tumour types.

Knowledge regarding symptom clusters may inform targeted interventions [48]. While the current study cannot attest to aetiology or causality within the found cluster symptoms, the main finding of interest for clinicians is the association between physical symptoms and psycho-social symptoms for patients facing severe symptoms. As such, it is advisable to assess whether a patient may profit from psycho-social help when suffering from (multiple) severe physical symptoms, in addition to treatment of the physical symptoms themselves.

In this dissertation, I focused on several facets of the use of PRMs in eHealth. Due to the nature of eHealth, PROMs are the instrument of choice to measure health of users, while PREMs are the instrument of choice to evaluate health care. Because PRMs are self-report instruments, they measure health in a more indirect manner than physical measurements. This creates a vulnerability in regards to validity, making research into measurement properties very important. What follows is a discussion of my findings where I will first focus on the validity and reliability of PRMs in eHealth, after which I will take a slightly broader view on validity and reliability of PRMs in health care and psychological science. Lastly, I will discuss the exciting possibilities for using the large datasets collected through routine use of PRMs.

Chapter 7

Discussion

7.1 Main findings

7.1.1 Measurement properties of PRMs

The first two aims of this dissertation were to investigate the measurement properties of various PROMs included in Oncokompas, as well as one widely used PREM in cancer care. In chapters 2 - 4 I presented the systematic reviews of measurement properties of two PROMs and one PREM, which are used in the eHealth application Oncokompas [8–11]. We found that the majority of measurement properties across the three PRMs were rated as either indeterminate (37.5%), or inconsistent (25%); with a little over one third rated as sufficient (37.5%). Furthermore, quality of evidence was mostly very low, low or moderate (81.8%), with a minority rated as high (18.2%).

As PRMs are often used in practice and research to inform on patient health and to evaluate health care, this requires further attention. In a broader systematic review on the 29 PRMs used in Oncokompas [8–11], we found that for many of these PROMs information was missing with respect to multiple measurement properties [51]. Table 7.1 shows an overview of which measurement properties were investigated and which were not. In most cases, we can see that certain measurement properties are well-established, most notably construct validity and internal consistency. As such, the main concern lies with suboptimal knowledge on reliability, measurement error, and responsiveness. Of these three reliability is particularly concerning. Without evidence of reliability, it raises the question whether a different outcome would have occurred if the PRM had been completed at a different point in time or in a slightly different setting. Furthermore, it is of note that a large proportion of studies which investigated structural validity, did so using outdated or subpar analyses. Structural validity determines whether subscales represent one construct, and is crucial for determining the validity of a PRM. As such, the research body on structural validity of PRMs is concerning.

Cross-cultural validity pertains to the question whether the validity of an instrument is the same across cultures. In most cases this refers to whether the structural validity is similar across cultures. For example, in the analysis of the structural validity of the eHealth Impact Questionnaire (chapter 5), we found it to be different in our three Dutch samples than in the original British sample for which it was developed and validated. However, this is not proof against cross-cultural validity, as an analysis is required to assess the measurement invariance in a dataset with both British and Dutch participants. Table 7.1 shows that a few instruments have been investigated on cross-cultural validity (most notable the EORTC instruments). For the remaining PROMs there is no evidence available, even though many have been translated to different languages and claim to be cross-culturally valid [241–243].

In the Intermezzo, I investigated the importance of more information on the measurement error of PRMs. Through use of simulations I showed the impact of measurement error on the results of research. With an increase in measurement error, I found an increase in bias on the parameter of interest, and a deflation of effect size. In the introduction, I described the effect that validity and reliability of PRMs may have on research into efficacy and cost-effectiveness of eHealth applications. The results of these simulations showed that when measurement error reaches 20% of the range of the instrument, the estimated effect sizes decreased dramatically, so that perceiving the effect seems nigh impossible. We currently have no indication as to what constitutes a small or large amount of measurement error. As such, we need thorough research into the measurement error of often-used PRMs for multiple reasons. Such research can allow the comparison of instruments to select the best-suited PRM for both research purposes and clinical practice. Furthermore, such research can provide a body of knowledge on the distribution of measurement error, helping to find an answer as to what constitutes relatively small or large amounts of measurement error. However, assessing measurement error is not enough, and we need thorough research in how to deal with it. While there are some proposals on how to deal with measurement error (e.g. using uninformed comparators [244]), such methods are focused on assessing how large the issue of measurement error is for a current sample, but not focused on how to adjust our data-analysis to correct for the issue.

It can be concluded that many measurement properties are either not examined or research shows inconsistent findings, and that the quality of research is subpar at best. These observations are in line with a recent editorial by Andrew Vickers [245], where he notes that validation studies of PRMs tend to conclude the PRM to be valid and reliable, regardless of the subpar methodology used. It should be noted that the COSMIN criteria were first published in 2010 and updated in 2018, while the majority of PRMs that are most commonly used were developed and published before 2010. It is unfair to fault original validation efforts for not using state-of-the-art validation techniques when these were not widely available at that time. Nevertheless, it should be noted that psychometrical methodologies that are currently recommended for use by the COSMIN, were developed before most PROMs were. For example, Confirmatory Factor Analysis (for structural validity) was developed in 1969 [246], the Intraclass Correlation Coefficient (for test-retest reliability) was developed in 1966 [247], and Omega (for internal consistency) was developed in 1970 [248]. Most of these methodologies were advocated to be used instead of the prevailing methodologies at that time by psychometricians in the 1990s [249–253], which is the era in which many of the most commonly used PRMs were developed and published. However, change in methodology takes time, and as such, these PRM developers should not be harshly

Table 7.1. Summary of available evidence (yes/no) on measurement properties of a selection of PROMs [51].

PROM	Structural Validity	Internal Consistency	Reliability	Measure-ment Error	Construct Validity	Criterion Validity	Responsive-ness	Cross-cultural Validity
Hospital Anxiety and Depression Scale	Yes	Yes	Yes	No	Yes	Yes	Yes	No
Cancer Worry Scale	Yes	Yes	No	No	Yes	Yes	No	No
SF-36: Cognitive function scale	No	Yes	No	No	Yes	No	Yes	No
Patiënt Specifieke Klachten (Dutch-specific)	No	No	No	No	No	No	No	No
Insomnia Severity Index	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
6-item Female Sexual Function Index	No	Yes	Yes	No	Yes	Yes	No	No
5-item International Index of Erectile Function	No	Yes	Yes	Yes	Yes	Yes	Yes	No
Body Image Scale	Yes	Yes	Yes	No	Yes	No	Yes	No
EORTC QLQ-C30	Yes	Yes	Yes	No	Yes	No	Yes	Yes
De Jong-Gierveld Loneliness Scale	Yes	Yes	Yes	No	Yes	No	No	No
Dyadic Adjustment Scale	Yes	Yes	Yes	No	Yes	No	No	No
Vragenlijst Gezinskenmerken Short-form (Dutch-specific)	Yes	Yes	Yes	No	Yes	No	No	No
EORTC IN-PATSAT32	Yes	Yes	Yes	No	Yes	No	No	No
Job Content Questionnaire	Yes	Yes	Yes	No	Yes	No	No	No
Vragenlijst Beleving en Beoordeling van de Arbeid (Dutch-specific)	Yes	Yes	No	No	No	No	No	No
Alcohol Five Shot	No	No	No	No	No	Yes	No	No
Perceived Stress Scale	Yes	Yes	Yes	No	No	No	No	No
Functional Assessment of Cancer Treatment - Endocrine Scale	No	Yes	Yes	No	Yes	No	Yes	No
Breast Impact of Treatment Scale	Yes	Yes	Yes	No	Yes	No	No	No
EORTC QLQ-BR23	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Quick Disabilities of the Arm, Shoulder, and Hand Questionnaire	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Breast Reconstruction Satisfaction Questionnaire	Yes	Yes	Yes	No	No	No	No	No
EORTC QLQ-CR29	Yes	Yes	Yes	No	Yes	No	Yes	Yes
Stoma Quality of Life Questionnaire	Yes	Yes	Yes	No	No	No	No	No
EORTC QLQ-H&N35	Yes	Yes	No	No	Yes	No	Yes	Yes
Shoulder Disability Questionnaire	No	Yes	Yes	No	Yes	Yes	Yes	No

judged. The larger issues lays with validation studies performed in more recent years. About as many validation studies were included in our systematic review after as before 2010. However, the methods used by these recent studies were often copied from the older original validation studies of the PRMs. There appears to be a disconnect between researchers performing validation studies, and the psychometricians who develop and advocate methodologies. This is especially notable, as the publications advocating methods (such as CFA, ICC, and Omega) have increased by an incredible amount in recent years.

The above is quite critical regarding the quality of research into the measurement properties of most PRMs, which might lead to the conclusion that data produced by PRMs can not be trusted. While I do not entirely discount this notion, I would advocate a different point of view. For many constructs measured by PRMs, we have no alternative which is better than a PRM. For example, the measurement of an attitude could be performed through a structured interview. However, the measurement properties of a structured interview are harder to investigate than those of a PRM, due to the non-standardized deviations in questions presented to patients. Another example would be the measurement of anxiety. While a structured interview conducted by a psychologist or psychiatrist is necessary for a diagnosis, a PROM has the distinct advantage of having standardized cut-off points to help identify those patients with an increased risk of an anxiety disorder, and of measuring the degree of anxiety as reported by a patient. As such, PRMs have a very distinct position in both research and clinical settings. What I wish to advocate is for users of PRMs, whether they are independent researchers or institutions, is to perform validation analyses on their PRM datasets or to publish these datasets so other researchers may do so. I will delve into further specifics in the section on 'Future research'.

In line with our second research aim, in chapter 5 I presented a study on the translation and validation of the Dutch version of the eHealth Impact Questionnaire, a PREM specifically designed to evaluate health care. Throughout this particular study we were able to follow the COSMIN criteria as updated in 2018, so that each relevant and feasible measurement property was investigated. And while a different structural validity was found compared to the original, through replication we found a consistent factor structure. With this validation study we offer a valid PREM to evaluate eHealth applications, which will hopefully help eHealth developers perform more efficient evaluation cycles. Furthermore, this study is an example of how to design a validation study using a methodology that would be rated well according to the COSMIN criteria.

7.1.2 Symptom clusters

The third aim was to investigate symptom clusters among cancer survivors. In chapter 6 we used the data of the PRMs used in Oncokompas to analyse the presence of symptoms in users to extract symptom clusters. We found an inherent difference on the co-occurrence of symptoms dependent on symptom severity. Among survivors with only high risk scores, the data showed a clustering of more connections between physical and psycho-social symptoms. Meanwhile, among survivors with moderate-to-high risk scores, we observed less connections in the clustering between physical and psycho-social symptoms. These findings are a valuable contribution to the literature of symptom clusters by using a broader measurement strategy than most previous studies [48], and by using a specific machine learning algorithm appropriate for the research question [234,237].

The investigation of symptom clusters may inform targeted interventions by identifying symptoms that may have an effect on other symptoms [48]. This is an intriguing and probably highly impactful line of research, which requires the investigation of symptom interactions, aetiology, and causality. In this line of research, we are currently in an explorative phase, where we are identifying those co-occurring symptoms. For the investigation of symptom interactions, very large datasets in which patients are measured across time are required. The use of electronic health records is a solid option, in particular because the use of digital PRMs to measure a large variety of symptoms is increasingly implemented in the Netherlands [4]. Such records contain data on a variety of symptoms over time, and as such the trajectory of one symptom may be associated to the trajectory of another symptom. Analysing such a research question will be complex, due to the large amount of possible symptom interactions as well as the requirement of analysing a large dataset of patients. This difficult task may be accomplished by forming hypotheses based on co-occurring symptoms identified in the explorative phase, which can then be tested using machine learning algorithms. It should be noted that a selection bias may be inherent to the use of digital PRMs, as 30% of cancer patients does not have access to them [254].

7.2 Limitations

There are a number of limitations to the research presented in this dissertation. In regards to chapters 2 - 4 one important measurement property was not investigated: content validity. Content validity is "the degree to which the content of a PROM is an adequate reflection of the construct to be measured" [38,39]. In short, content validity is judged to be good when the items of the PRM are "relevant, comprehensive, and comprehensible with respect to the construct of interest and study population" [38]. Because the judgement of content validity is based on a subjective judgement of

reviewers, where specific expertise on the construct that is measured is paramount, it was beyond the scope of this dissertation. However, it should be noted that all other measurement properties are only relevant if there is evidence of good content validity as well.

Another limitation in regards to chapters 2 - 4, and the results as presented in Table 7.1, is the use of a precise rather than a sensitive search filter regarding measurement properties. While the sensitivity of the used search filter was 93% [61], there is a possibility that validation studies were missed. As the search was focussed on many PRMs, the use of the precise filter was a pragmatic choice to limit search hits. To alleviate this limitation we performed manual searches for the EORTC IN-PATSAT32, IIEF, and FSFI and found no missing records.

In regards to chapter 5, the limitations mostly refer to the findings themselves, as there were a number of under-performing measurement properties of the modified factor structure. In particular, test-retest reliability was below acceptable (ICC < .70) for two subscales, and some correlations for the test of convergent validity were small (r < .30). The former requires further study in new samples, while the latter may be explained by subpar a priori hypotheses in regard to the instruments used for convergent validity. Regardless, convergent validity needs to be more thoroughly studied in future validation endeavours.

Lastly, for chapter 6, there are limitations in regards to the data-analysis strategy. The analysis was focused on the extraction of smaller clusters. Due to this, there was a high count of data points deemed as noise. This indicates that there may be other (likely larger) clusters that could be extracted using a different analysis strategy. Furthermore, for users that filled in Oncokompas multiple times, we chose a random data row. This method ensures we do not increase bias in our data set. However, a biased data set (e.g. selecting the data row which includes most moderate-to-high risk scores) could theoretically produce a differently informed result. Finally, the results of this study are possibly biased: we know that we do not reach approximately 30% of cancer survivors with online tools as Oncokompas, which is, among others, related to female gender, older age, and lower health literacy [254].

While this dissertation identified a number of issues with regard to research towards measurement properties of PRMs and also offers suggestions towards improving such research, it would have been of added value to implement these suggestions specifically for the PRMs we investigated. While we implemented a thorough validation of the eHIQ (chapter 5), there were no immediate issues with the previous validation of the eHIQ. Instead of merely suggesting how validation of the IIEF, FSFI, and EORTC IN-

PATSAT32 could be improved, new validation studies of these PRMs would have been a strong start towards stronger evidence of their measurement properties.

7.3 Implications of findings

7.3.1 Measurement properties

The issues discussed previously, in regard to measurement properties, are not isolated to the use of PRMs in eHealth and for the evaluation of eHealth applications. The three instruments that we evaluated (chapter 2 - 4), the two instruments on which reviews were published elsewhere [58,59], and the remaining PRMs we investigated in less depth [51], are all used outside of eHealth and eHealth research as well. As discussed in the introduction of this dissertation, PRMs are used in routine health care and in research evaluating health care [4,5]. However, PRMs are also used to measure psychological constructs in the broader field of psychological and behavioural science. Most of the PRMs we have investigated do not target psychological constructs, but the question arises whether the same issues we have identified regarding measurement properties arise in the PRMs used in psychological science as well. This is of interest, since in recent years there has been a focus on a "reproducibility crisis" in the psychological field, with many effects not reproducing in new studies [255,256]. While a large range of factors influencing and ways to solve this phenomenon have been proposed [257–260], methods of measuring constructs has been under-represented as an underlying factor. It is conceivable that if the measurement instruments used are not as valid and reliable as we assume them to be, the results they produce may not be precise enough to enable replication. As such, it is of importance to further investigate the measurement properties of PRMs used in psychological science.

7.3.2 Routinely collected data

Chapter 6 illustrates the possibility of using PRM data to investigate relevant theoretical research questions. With the increase of eHealth usage [7], and PRMs being adopted by Dutch hospitals [4] and Dutch health care insurers [5] to implement and focus on value-based health care, large datasets of PRM responses are gathered. These datasets provide the means to efficiently investigate certain research questions, that would otherwise require a lot of resources to investigate [42]. For example, the investigation of symptom clusters previously required sampling hundreds of patients in a hospital, using very specific short measurement tools [46,47] as to not overburden the patients. The use of short measurement tools restricts the amount of symptoms that could be measured. Meanwhile, data collected through eHealth applications and in electronic patient files does not expend any extra resources or is anymore taxing of patients, than their regular use of the application.

7.4 Future research

More thorough research is needed in the investigation of measurement properties of PRMs, but running validation studies is costly and requires resources that many research teams do not have to spare. In recent years, the movement towards open data has gathered momentum, with multiple platforms for researchers to share their data, such as the Dataverse [261,262], LinkedScience [263] and the Open Science Framework [264,265]. Open data proponents are usually focused on the possibility of reproducing and checking scientific results [256,266], but open data sets could be used to perform certain validation analyses. Datasets that do not originate from a study aimed at evaluating measurement properties, are by default limited in the information that they can provide. However, such datasets can easily be analysed to inform on the structural validity of the measurement instrument, granted that the sample size is large enough. As previously mentioned, structural validity is a measurement property in which our knowledge is often lacking, while tests of unidimensionality are paramount for valid use of an instrument. A benefit of using open datasets is that they can be very varied in the populations they investigate. Running validation analyses on multiple such datasets will help establish a more generalizable observation of measurement properties.

To be more concrete on the possibilities of analyses of such datasets, I refer back to Samples 2 and 3 which were used to validate the eHealth Impact Questionnaire in chapter 5. These samples were not gathered specifically for the use of validation, but were collected as part of a RCT investigating the efficacy of Oncokompas (sample 2) [8], or as part of a pilot study of an eHealth app providing health information regarding pre- and post-operative care (sample 3). We analysed these datasets using Confirmatory Factor Analyses, Exploratory Factor Analyses, and we fitted Graded Response Models to investigate structural validity. With this analysis approach we were able to confirm a factor structure which was consistent across Dutch samples. If such analyses were routinely performed on existing data sets, the knowledge gap could be quickly filled regarding structural validity of the measurement instruments that we use. By combining existing datasets on which structural validity can be tested, cross-cultural validity can be tested, further filling our current knowledge gap regarding validity of our measurement instruments.

The previous paragraphs focused on open datasets, but the previously discussed routinely collected data could serve the same purpose. There are two distinct advantages of using these datasets in comparison to open datasets: these particular datasets are likely to be large enough for validity analyses, and are based on PRMs used in routine health care. By investigating measurement properties on data that is routinely collected, we could create an evaluation-loop where PRMs used in eHealth and health care could be updated and improved over time. The use of routinely collected data for research purposes carries

certain privacy risks. Therefore, steps need to be taken to protect the privacy of patients. Anonymizing patient data is problematic to perform, due to which other steps need to be taken. For example, all analyses could be run on the same server architecture that the data is collected on, so that the data never leaves its privacy-certified environment.

While the solution proposed in the previous paragraphs could cover our gap in knowledge in regards to structural validity of our instruments, we still lack knowledge on test-retest reliability, measurement error, and responsiveness. Assessing these measurement properties requires a specific methodological set-up. To assess test-retest reliability and measurement error, a design is required where the instrument is used (at least) twice in a short period of time in a similar setting. This design was used in sample 1 of chapter 5. This creates a barrier to entry: many researchers do not have the time or resources to set up a specific validation study such as this. However, with recent technological advancements, such studies can actually be run with a relatively low resource cost. Crowdsourcing is "the distribution of tasks to large groups of individuals via a flexible open call" [267]. Internet platforms that implement crowdsourcing have popped up in recent years, most notably Amazon's Mechanical Turk [268] which has been assessed as an appropriate platform to perform experimental [269], and clinical [267] psychological research. Some other similar platforms are Clickworker [270] and MicroWorkers [271]. The use of crowdsourcing platforms creates a low-cost method of getting our instruments used by either a very diverse or very specific sample of participants, allowing researchers to conduct validation studies while expending less time and resources.

Our issues do not end with conducting more methodologically sound validation analyses and studies. After we have gathered more knowledge on measurement properties of PRMs, this knowledge also needs to be conveyed to researchers and clinicians. With dozens of PRMs aimed to measure the same construct, it can be an arduous task to choose the most appropriate measurement instrument. Due to this, often one or two PRMs become the standard in a certain field, usually not due to the merits of its measurement properties, but due to convention created by prominent leaders of their field. It is plausible that most criticisms of certain PRMs only reach a minority of researchers and clinicians. Mirroring open data platforms [261–265] - as well as being in line with the movement towards open science by the European Union [272] - a platform could be devised where researchers who have performed validation analyses could upload their results, preferably including the dataset itself. By using machine readable formats [273], an automatic qualitative aggregation of measurement properties could be created. This aggregation could then be used by researchers and clinicians to help with their decision on which PRM is most fit to be used for their purposes. An extra benefit of such a platform would be the option to combine datasets for meta-validation analyses (e.g. the aforementioned tests of measurement invariance).

7.5 Conclusion

In this dissertation I explored the measurement properties of various PRMs that are used internationally in clinical and research settings. Based on the COSMIN criteria, I conclude that many PRMs require more thorough research to provide evidence on certain measurement properties. The available evidence is often based on studies using outdated methodologies and seem to be modelled after validation studies performed between 1990 and 2010. Through the use of the COSMIN criteria, more thorough validation studies may be designed. Using the COSMIN criteria as a guideline, the eHealth Impact Questionnaire was translated and validated for the Dutch population of eHealth users.

In the meantime, the use of routinely collected large datasets of PRM responses provide interesting opportunities for research. An example is the study on symptom clusters among cancer survivors, using machine learning techniques on a large data set of PRM responses from Oncokompas users. The results indicated an inherent difference on the co-occurrence of symptoms dependent on symptom severity. Cancer survivors with high risk of non-wellbeing for certain symptoms showed more connections between physical and psycho-social symptoms than among survivors with moderate-to-high risk of non-wellbeing.

Since validation studies require a lot of resources and time to carry out, certain measurement properties could be analysed by running validation analyses on open datasets or routinely collected data. Currently, routinely collected data is mostly used in clinical care but underused for scientific endeavours. Beyond validation analyses many interesting hypotheses could be tested in these large datasets, and the scientific community and health care system may be aided through these innovative endeavours.


E

Epilogue

Summary

Patient Reported Measures (PRMs) are instruments completed by patients to measure various constructs. PRMs can be subdivided into two main categories: Patient Report Outcome Measures (PROMs) measure health-related quality of life and symptoms of the individual patient, while Patient Reported Experience Measures (PREMs) evaluate the quality of health care from the perspective of the patient. In this dissertation, the focus lies on PROMs and PREMs which are used in eHealth which pertains the provision of health care services through digital media. Oncokompas is an eHealth self-management application that supports Dutch cancer survivors in finding and obtaining optimal supportive care, adjusted to their personal health status and preferences. To provide personally adjusted advice, Oncokompas uses 29 widely used PRMs (besides several newly developed PRMs). The first aim of this dissertation is to investigate the measurement properties of various PRMs included in Oncokompas.

Measurement properties refer to the validity and reliability of a measurement instrument, which are crucial to determine whether the measurement instrument is capable of being used in practice. Validity is "the degree to which a measurement instrument measures the construct(s) purport to be measure", and reliability is "the degree to which the measurement is free from measurement error". Validity and reliability can be broken down into subcategories (also called measurement properties). The COnsensus-based Standards for the selection of health status Measurement INstruments (COSMIN) taxonomy and COSMIN guidelines provide a framework for discourse and interpretation of these different subcategories, specifically for PRMs. In order to investigate the measurement properties of the 29 existing PROMs and one PREM used in Oncokompas, we performed a systematic review using the COSMIN guidelines. While discussing all of the results of this systematic review is beyond the scope of this dissertation, in this dissertation, we delve deeper into the measurement properties of two PROMs that aim to assess sexuality (the International Index of Erectile Function in chapter 2, and the Female Sexual Function Index in chapter 3), and one PREM that aims to measure satisfaction with in-patient cancer care (the EORTC IN-PATSAT32 in chapter 4).

The evaluation of eHealth applications presents very specific issues. Scientific evaluation using randomized controlled trials or in-depth evaluation through user experience interviews take a lot of time and resources. Meanwhile, the development of eHealth applications is usually rapid, leading to a state of "playing catch-up" for eHealth developers. The eHealth Impact Questionnaire (eHIQ) is a PREM designed to measure a users attitude towards eHealth. The second aim of this dissertation is to translate and validate the eHIQ for the Dutch population of eHealth users (chapter 5).

The use of validated and reliable PRMs in health care creates exciting possibilities. As mentioned, the use of PRMs has been promoted in routine health care in the Netherlands. PRMs are filled in by a patient at various stages of treatment, nowadays often through use of an eHealth application (e.g. a PRM presented through a website). Through these digitized PRMs an enormous amount of data is gathered. These big data sets can be used to explore theoretical questions that thus far could not be investigated on such a large scale. The third and last aim of this dissertation is to investigate symptom clusters among cancer survivors using the large dataset collected by Oncokompas to investigate symptom clusters among cancer survivors (chapter 6.

The International Index of Erectile Function (IIEF) is a PROM to evaluate erectile dysfunction and other sexual problems in males. We performed a systematic review of the measurement properties of the IIEF-15 and the IIEF-5. A systematic search of scientific literature up to April 2018 was performed. Data were extracted, and analysed according to COSMIN guidelines for structural validity, internal consistency, reliability, measurement error, hypothesis testing for construct validity and responsiveness. Evidence of measurement properties was categorized into sufficient, insufficient, inconsistent, or indeterminate, and quality of evidence as very high, high, moderate, or low. The main outcome measure was the evidence of a measurement property, and the quality of evidence based on the COSMIN guidelines. Forty studies were included. The evidence for criterion validity (of the Erectile Function subscale), and responsiveness of the IIEF-15 was sufficient (high quality), but inconsistent (moderate quality) for structural validity, internal consistency, construct validity, and test-retest reliability. Evidence for structural validity, test-retest reliability, construct validity, and criterion validity of the IIEF-5 was sufficient (moderate quality), but indeterminate for internal consistency, measurement error and responsiveness. Lack of evidence for and evidence not supporting some of the measurement properties of the IIEF-15 and IIEF-5, shows the importance of further research on the validity of these questionnaires in clinical research and clinical practice. A strength of the review was the use of pre-defined guidelines (COSMIN). A limitation of the review was the use of a precise rather than a sensitive search filter regarding measurement properties to identify studies to be included. The IIEF requires more research on structural validity (IIEF-15), internal consistency (IIEF-15 and IIEF-5), construct validity (IIEF-15), measurement error (IIEF-15 and IIEF-5), and responsiveness (IIEF-5). The most pressing matter for future research is determining the unidimensionality of the IIEF-5, and the exact factor structure of the IIEF-15.

The Female Sexual Function Index (FSFI) is a PROM measuring Female Sexual Dysfunction (FSD). The FSFI-19 was developed with six theoretical subscales in 2000. In 2010, a shortened version became available (FSFI-6). We performed a systematic review to investigate the measurement properties of the FSFI-19 and FSFI-6. A

systematic search was performed of Embase, Medline, and Web of Science for studies that investigated measurement properties of the FSFI-19 or FSFI-6 up to April 2018. Data were extracted, and analyzed according to COSMIN guidelines. Evidence was categorized into sufficient, insufficient, inconsistent, or indeterminate, and quality of evidence as very high, high, moderate, or low. The main outcome measure was the evidence of a measurement property, and the quality of evidence based on the COSMIN guidelines. Eighty-three studies were included. Concerning the FSFI-19, the evidence for internal consistency was sufficient and of moderate quality. The evidence for reliability was sufficient but of low quality. The evidence for criterion validity was sufficient and of high quality. The evidence for structural validity was inconsistent of low quality. The evidence for construct validity was inconsistent of moderate quality. Concerning the FSFI-6, the evidence for criterion validity was rated as sufficient of moderate quality. The evidence for internal consistency was rated as indeterminate. The evidence for reliability was inconsistent of low quality. The evidence for construct validity was inconsistent of very low quality. No information was available on structural validity of the FSFI-6, and measurement error, responsiveness, and cross-cultural validity of both FSFI-6 and FSFI-19. Conflicting and lack of evidence for some of the measurement properties of the FSFI-19 and FSFI-6, indicates the importance of further research on the validity of these PROMs. We advise researchers whom use the FSFI-19 to perform confirmatory factor analyses and report the factor structure found in their sample. Regardless of these concerns, the FSFI-19 and FSFI-6 have strong criterion validity. Pragmatically, they are good screening tools for the current definition of FSD. A strength of the review was the use of pre-defined guidelines. A limitation was the use of a precise rather than a sensitive search filter. The FSFI requires more research on structural validity (FSFI-19 and FSFI-6), reliability (FSFI-6), construct validity (FSFI-19), measurement error (FSFI-19 and FSFI-6), and responsiveness (FSFI-19 and FSFI-6). Further corroboration of measurement invariance (both across cultures and across subpopulations) in the factor structure of the FSFI-19 is necessary, as well as tests for the unidimensionality of the FSFI-6.

The EORTC IN-PATSAT32 is a patient reported outcome measure (PROM) to assess cancer patients' satisfaction with in-patient health care. We investigated whether the initial good measurement properties of the IN-PATSAT32 were confirmed in new studies. Within the scope of a larger systematic review study (Prospero ID 42017057237), a systematic search was performed of Embase, Medline, PsycINFO, and Web of Science for studies that investigated measurement properties of the IN-PATSAT32 up to July 2017. Study quality was assessed, data were extracted, and synthesized according to the COSMIN guidelines. Nine studies were included in this review. The evidence on reliability and construct validity were rated as sufficient and of the quality of the evidence as moderate. The evidence on structural validity was rated as insufficient and

of low quality. The evidence on internal consistency was indeterminate. Measurement error, responsiveness, criterion validity, and cross-cultural validity were not reported in the included studies. Measurement error could be calculated for two studies, and was judged indeterminate. In summary, the IN-PATSAT32 performs as expected with respect to reliability and construct validity. No firm conclusions can be made yet whether the IN-PATSAT32 also performs as well with respect to structural validity and internal consistency. Further research on these measurement properties of the PROM is therefore needed as well as on measurement error, responsiveness, criterion validity, and cross-cultural validity. For future validation studies, it is recommended to take the COSMIN methodology into account.

Measurement Error represents the minimum amount of change measured by a measurement tool, of which we can be sure is not an artefact of systematic error. In a large-scale systematic review, we found that only 4.14% of validation articles reported on measurement error, and measurement error could be calculated for another 3.82% of articles. To illustrate the implications measurement error has on clinical research, a simulation study was conducted. Simulations were run on a hypothetical randomized controlled trial for the treatment of depression as measured by the Beck Depression Inventory-II. Baseline values and a decrease over time of depressive symptoms for untreated depression (control condition) were extracted from literature. The Minimal Clinically Important Difference (MCID) was used as a measure of effect size for the further decrease over time of the treatment condition. Three parameters were systematically varied across simulations: sample size (250 / 500 / 750), effect size (0*MCID / 1*MCID / 2*MCID / 3*MCID), and measurement error (0% / 10% / 20% / 30% / 40%). Each parameter combination was simulated 5000 times. The relative bias is the bias of the coefficient of interest. The relative bias became more biased from near zero (with no measurement error) to -0.5 (with 30% and 40% measurement error). Furthermore, higher effect sizes showed more relative bias. ETA Squared is a measure of effect size. The ETA Squared ranges from 0 to 0.525 when there is 0% measurement error, dependent on the effect size parameter. Every ETA squared drifted further towards zero with more added measurement error. The results of the simulation showed an increase in bias with the addition of more measurement error. Furthermore, this effect seemed to be stronger for higher effect sizes. The result of this bias is a decrease of effect size, which is especially dramatic upwards of 20% measurement error. It appears that measurement error affects power to detect a true effect.

The eHealth Impact Questionnaire (eHIQ) provides a standardized method to measure attitudes of electronic health (eHealth) users towards eHealth. It has previously been validated in a population of eHealth users in the United Kingdom, and consists of 2 parts and 5 subscales. Part 1 measures attitudes toward eHealth in general and consists

of the subscales *Attitudes towards online health information* (5 items), and *Attitudes towards sharing health experiences online* (6 items). Part 2 measures the attitude towards a particular eHealth application and consists of the subscales *Confidence and identification* (9 items), *Information and presentation* (8 items), and *Understand and motivation* (9 items). The eHIQ was translated and validated in accordance with the COSMIN criteria. The validation comprised 3 study samples with a total of 1287 participants. Structural validity was assessed using confirmatory factor analyses and exploratory factor analyses (EFAs; all 3 samples). Internal consistency was assessed using hierarchical omega (all 3 samples). Test-retest reliability was assessed after 2 weeks, using two-way intraclass correlation coefficients (sample 1). Measurement error was assessed by calculating the smallest detectable change (sample 1). Convergent and divergent validity were assessed using correlations with the remaining measures (all 3 samples). A graded response model was fit and item information curves were plotted to describe the information provided by items across item trait levels (all 3 samples). The original factor structure showed a bad fit in all 3 study samples. EFAs showed a good fit for a modified factor structure in the first study sample. This factor structure was subsequently tested in sample 2 and 3, and showed acceptable to good fits. Internal consistency, test-retest reliability, convergent validity, and divergent validity were acceptable to good for both the original as the modified factor structure, except for test-retest reliability of one of the original subscales, and the 2 derivative subscales in the modified factor structure. The graded response model showed that some items underperformed in both the original and modified factor structure. The Dutch version of the eHIQ (eHIQ-NL) shows a different factor structure compared with the original English version. Part 1 of the eHIQ-NL consists of 3 subscales: *Attitudes towards online health information* (5 items), *Comfort with sharing health experiences online* (3 items), and *Usefulness of sharing health experiences online* (3 items). Part 2 of the eHIQ-NL consists of three subscales: *Motivation and confidence to act* (10 items), *Information and presentation* (13 items), and *Identification* (3 items).

Knowledge regarding symptom clusters may inform targeted interventions. We investigated symptoms clusters among cancer survivors, using machine learning techniques on a large data set. Data were used of cancer survivors who used the fully automated online application 'Oncokompas'. Oncokompas supports survivors in their self-management by 1) monitoring their symptoms through PROMs; and 2) providing tailored feedback on their scores with a personalized overview of supportive care options, aiming to reduce symptoms burden and improve health-related quality of life. In the present study, data on 26 generic symptoms (physical and psychosocial) were used. Results of the PROM of each symptom are presented to the user as a no well-being risk, moderate well-being risk, or high well-being risk score. Data of 1032 cancer survivors were analysed using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) on high risk scores and moderate-to-high risk scores separately.

When analysing the high risk scores, seven clusters were extracted: one main cluster which contained most frequently occurring physical and psychosocial symptoms, and six subclusters with different combinations of these symptoms. When analysing moderate-to-high risk scores, three clusters were extracted: two main clusters were identified, which separated physical symptoms (and their consequences) and psychosocial symptoms, and one subcluster with only body weight issues. There appears to be an inherent difference on the co-occurrence of symptoms dependent on symptom severity. Among survivors with high risk scores, the data showed a clustering of more connections between physical and psychosocial symptoms in separate subclusters. Among survivors with moderate-to-high risk scores, we observed less connections in the clustering between physical and psychosocial symptoms.

Across chapters 2, 3 and, 4 we investigated the measurement properties of three PRMs. We found that the majority of measurement properties across the three PRMs were rated as either indeterminate (37.5%), or inconsistent (25%); with a little over one third rated as sufficient (37.5%). We also found that the quality of evidence was mostly very low, low or moderate (81.8%), with a minority rated as high (18.2%). Furthermore, in a broader systematic review on the 29 PRMs used in Oncokompas, we found that for many of these PRMs information was missing with respect multiple measurement properties. This is concerning, as PRMs are often used in practice and research to inform on patient health and to evaluate health care. In particular, more research is necessary on reliability, measurement error, responsiveness, and cross-cultural validity. The validation study performed of the eHIQ-NL (chapter 5) serves as an example of how a validation study can be performed which would rate well on the COSMIN guidelines.

Chapter 6 illustrates the possibility of using PRM data to investigate relevant theoretical research questions. With the increase of eHealth usage, and PRMs being adopted by Dutch hospitals and Dutch health care insurers to implement and focus on value-based health care, large datasets of PRM responses are gathered. These datasets can be used to investigate research questions, that would otherwise require lots of resources to investigate. The investigation of symptom clusters is one such research question, and routinely collected data could be used to further this line of research. Routinely collected data could also be used for validation analyses, most notably in the investigation of structural validity for which evidence is often lacking. Open datasets published on platforms such as Dataverse, LinkedScience, and the Open Science Framework could be used in similar fashion. Investigation into test-retest reliability, measurement error, and responsiveness requires a more specific methodological design. To reduce the resources needed for such studies, crowdsourcing may be used. Results of such research also needs to be conveyed more appropriately towards clinicians and researchers that actually use

the measurement instrument. Mirroring open data platforms - as well as being in line with the the movement towards open science by the European Union - a platform could be devised where researchers who have performed validation analyses could upload their results, preferably including the dataset itself. By using machine readable formats, an automatic qualitative aggregation of measurement properties could be created.

Nederlandse Samenvatting

Patient Reported Measures (PRMs), door de patiënt ingevulde instrumenten, worden ingezet om verscheidene constructen te meten. PRMs kunnen verdeeld worden in twee hoofdcategorieën: Patient Reported Outcome Measures (PROMs), gebruikt om Health Related Quality of Life (HRQoL) en symptomen van de individuele patiënt te meten, terwijl Patient Reported Experience Measures (PREMs) de kwaliteit van de gezondheidszorg evalueren vanuit het perspectief van de patiënt. In dit proefschrift ligt de focus op PROMs en PREMs die gebruikt worden in eHealth, het aanbieden van gezondheidszorg door middel van digitale media. Het Oncokompas is een eHealth zelfmanagement applicatie om kankerpatiënten te ondersteunen in het vinden en verkrijgen van optimale ondersteunende zorg, aangepast aan hun persoonlijke gezondheid en voorkeuren. Om persoonlijk advies te kunnen geven maakt het Oncokompas gebruik van 29 veelgebruikte PRMs (naast enkele nieuw ontwikkelde PRMs). Het eerste doel van dit proefschrift is het onderzoeken van de psychometrische eigenschappen van verscheidende PRMs opgenomen in het Oncokompas.

Psychometrische eigenschappen verwijzen naar de validiteit en betrouwbaarheid van een meetinstrument en zijn cruciaal om te bepalen of een meetinstrument in de praktijk gebruikt kan worden. Validiteit is “de mate waarin een instrument het construct meet waarvan wordt beweerd dat het wordt gemeten” en betrouwbaarheid is “de mate waarin het instrument vrij is van meetfout”. Validiteit en betrouwbaarheid kunnen opgedeeld worden in subcategorieën (ook wel psychometrische eigenschappen genoemd). De COnsensus-based Standards for the selection of health status Measurement INstruments (COSMIN) taxonomie en COSMIN richtlijnen verschaffen een kader voor het bespreken en interpreteren van verschillende van deze subcategorieën, specifiek voor PRMs. Om de psychometrische eigenschappen van 29 bestaande PROMs en één PREM gebruikt in het Oncokompas te onderzoeken is een systematische review uitgevoerd aan de hand van de COSMIN richtlijnen. Hoewel de volledige bespreking van de resultaten van deze systematische review buiten de reikwijdte van dit proefschrift ligt, zal ik me in dit proefschrift verdiepen in de psychometrische eigenschappen van twee PROMs gericht op seksualiteit (de International Index of Erectile Function in hoofdstuk 2 en de Female Sexual Function Index in hoofdstuk 3), en één PREM gericht op het meten van tevredenheid met intramurale zorg in kankerpatiënten (de EORTC IN-PATSAT32 in hoofdstuk 4).

Bij de evaluatie van eHealth applicaties komen erg specifieke problemen voor. Wetenschappelijke evaluaties door middel van randomized controlled trials of diepte interviews gericht op de ervaringen van patiënten kosten veel tijd en middelen. Ondertussen staat de ontwikkeling van eHealth applicaties niet stil, resulterend in een continue uitdaging om bij te blijven voor eHealth ontwikkelaars. De eHealth Impact

Questionnaire (eHIQ) is een PREM gericht op het meten van de houding van patiënten ten opzichte van eHealth. Het tweede doel van dit proefschrift is het vertalen en valideren van de eHIQ voor de Nederlandse populatie van eHealth gebruikers (hoofdstuk 5).

Het gebruik van gevalideerde en betrouwbare PRMs in de gezondheidszorg creëert opwindende mogelijkheden. Zoals hierboven genoemd, wordt het gebruik van PRMs aangemoedigd in de reguliere gezondheidszorg in Nederlands. PRMs worden ingevuld door patiënten in verscheidene behandelstadia, in de huidige tijd dikwijls doormiddel van een eHealth applicatie (zoals een PRM gepresenteerd op een website). Door middel van deze gedigitaliseerde PRMs wordt een enorme hoeveelheid data verzameld. Deze grote datasets kunnen gebruikt worden om theoretische vragen te verkennen, die tot op heden niet op een dergelijk grote schaal onderzocht konden worden. Het derde en laatste doel van dit proefschrift is het onderzoeken van symptoomclusters binnen kankerpatiënten door het gebruik van een grote dataset verzameld binnen het Oncokompas (hoofdstuk 6).

De International Index of Erectile Function (IIEF) is een PROM om erectieproblemen en andere seksuele problemen bij mannen te evalueren. We hebben een systematische review uitgevoerd naar de psychometrische eigenschappen van de IIEF-15 en de IIEF-5. Een systematische doorzoeking van de wetenschappelijke literatuur tot en met april 2018 werd uitgevoerd. Data werd geëxtraheerd en geanalyseerd volgens de COSMIN richtlijnen voor structurele validiteit, interne consistentie, betrouwbaarheid, meetfout, hypothese testen voor construct validiteit en responsiviteit. Bewijs voor psychometrische eigenschappen werd gecategoriseerd in voldoende, onvoldoende, inconsistent, of niet-bepaalbaar. De kwaliteit van bewijs was erg hoog, hoog, gemiddeld of laag. Veertig studies werden geïncludeerd. Het bewijs voor criterium validiteit (van de Erectile Function subschaal), en responsiviteit van de IIEF-15 was voldoende (hoge kwaliteit), maar inconsistent (gemiddelde kwaliteit) voor structurele validiteit, interne consistentie, construct validiteit, en test-hertest betrouwbaarheid. Bewijs voor structurele validiteit, test-hertest betrouwbaarheid, construct validiteit en criterium validiteit van de IIEF-5 was voldoende (gemiddelde kwaliteit), maar niet bepaaldbaar voor interne consistentie, meetfout en responsiviteit. De afwezigheid van bewijs voor, en de aanwezigheid van bewijs tegen een aantal psychometrische eigenschappen van de IIEF-15 en IIEF-5 benadrukken het belang van verder onderzoek naar de validiteit van dergelijke vragenlijsten in klinisch onderzoek en de klinische praktijk. Een kracht van de review was het gebruik van vooraf-gedefinieerde richtlijnen (COSMIN). Een beperking van de review was het gebruik van een precieze, in plaats van een sensitieve, zoekfilter met betrekking op psychometrische eigenschappen voor het identificeren van studies. De IIEF vereist verder onderzoek naar structurele validiteit (IIEF-15), interne consistentie (IIEF-15 en IIEF-5), construct validiteit (IIEF-15), meetfout (IIEF-15 en IIEF-5) en

responsiviteit (IIEF-5). De meest urgente kwestie voor vervolgonderzoek is de bepaling van unidimensionaliteit van de IIEF-5 en de exacte factor structuur van de IIEF-15.

De Female Sexual Function Index (FSFI) is een PROM om seksuele stoornissen bij vrouwen te meten. DE FSFI-19 werd in 2000 ontwikkeld met zes theoretische subschalen. In 2010 kwam een verkorte versie beschikbaar (FSFI-6). Een systematische doorzoeking van Embase, Medline en Web of Science werd uitgevoerd naar studies gericht op psychometrische eigenschappen van de FSFI-19 en FSFI-6 tot en met april 2018. Data werd geëxtraheerd en geanalyseerd volgens de COSMIN richtlijnen. Bewijs werd gecategoriseerd in voldoende, onvoldoende, inconsistent, of niet-bepaalbaar, en de kwaliteit van het bewijs als erg hoog, hoog, gemiddeld of laag. De belangrijkste uitkomstmaat was bewijs voor een psychometrische eigenschap, en de kwaliteit van dit bewijs volgens de COSMIN richtlijnen. Drieëntachtig studies werden geïncludeerd. Met betrekking tot de FSFI-19, was het bewijs voor interne consistentie voldoende en van gemiddelde kwaliteit. Het bewijs voor betrouwbaarheid was voldoende maar van lage kwaliteit. Het bewijs voor criterium validiteit was voldoende en van hoge kwaliteit. Het bewijs voor structurele validiteit was inconsistent en van lage kwaliteit. Het bewijs voor construct validiteit was inconsistent en van gemiddelde kwaliteit. Met betrekking tot de FSFI-6, werd het bewijs voor criterium validiteit beoordeeld als voldoende en van gemiddelde kwaliteit. Het bewijs voor interne consistentie werd beoordeeld als niet bepaalbaar. Het bewijs voor betrouwbaarheid was inconsistent en van lage kwaliteit. Het bewijs voor construct validiteit was inconsistent en van erg lage kwaliteit. Er was geen informatie beschikbaar met betrekking tot structurele validiteit van de FSFI-6 en meetfout, responsiviteit en cross-culturele validiteit van zowel de FSFI-6 en FSFI-19. Tegenstrijdig bewijs en de afwezigheid van bewijs voor een aantal psychometrische eigenschappen van de FSFI-19 en FSFI-6 benadrukken het belang van verder onderzoek naar de validiteit van dergelijke PROMs. We adviseren onderzoekers die gebruikmaken van FSFI-19 om confirmatieve factor analyses uit te voeren en de gevonden factorstructuur in hun steekproef te rapporteren. Los van deze zorgen hebben de FSFI-19 en de FSFI-6 sterke criterium validiteit. Pragmatisch gezien zijn het goede screeningsinstrumenten voor de huidige definitie van seksuele stoornissen bij vrouwen. Een kracht van de review was het gebruik van vooraf-gedefinieerde richtlijnen (COSMIN). Een beperking van de review was het gebruik van een precieze, in plaats van een sensitieve, zoekfilter. De FSFI vereist verder onderzoek naar structurele validiteit (FSFI-19 en FSFI-6), betrouwbaarheid (FSFI-6), construct validiteit (FSFI-19), meetfout (FSFI-19 en FSFI-6) en responsiviteit (FSFI-19 en FSFI-6). Verdere bevestiging van de meetinvariantie (zowel tussen culturen als tussen subpopulaties) in de factorstructuur van de FSFI-19 is noodzakelijk, net als tests voor de unidimensionaliteit van de FSFI-6.

The EORTC IN-PATSAT32 is een PROM om de tevredenheid met intramurale zorg van kankerpatiënten te meten. We onderzochten of de eerste goede psychometrische eigenschappen van de IN-PATSAT32 bevestigd werden in nieuwe onderzoeken. Binnen een grotere systematische review studie (Prospero ID 42017057237), werd een systematische doorzoeking van Embase, Medline, PsycINFO en Web of Science uitgevoerd naar studies gericht op de psychometrische eigenschappen van de IN-PATSAT32 tot juli 2017. Kwaliteit van de studies werd bepaald, data geëxtraheerd en samengevat volgens de COSMIN methodologie. Negen studies werden geïncludeerd in de review. Het bewijs met betrekking tot betrouwbaarheid en construct validiteit werd beoordeeld als voldoende en de kwaliteit als gemiddeld. Het bewijs voor structurele validiteit werd beoordeeld als onvoldoende en van lage kwaliteit. Het bewijs voor interne consistentie was niet bepaalbaar. Meetfout, responsiviteit, criterium validiteit en cross-culturele validiteit werden niet gerapporteerd in de geïncludeerde studies. Meetfout kon voor twee studies berekent worden en werd beoordeeld als niet bepaalbaar. Samenvattend presteert de IN-PATSAT32 zoals verwacht met betrekking tot betrouwbaarheid en construct validiteit. Er kunnen op dit moment geen harde conclusies getrokken worden over de prestatie van de IN-PATSAT32 met betrekking tot structurele validiteit en interne consistentie. Verder onderzoek is noodzakelijk naar de meetfout, responsiviteit, criterium validiteit en cross-culturele validiteit van deze PROM. Voor toekomstige validiteitsstudies, is het raadzaam om de COSMIN methodologie in acht te nemen.

Meetfout vertegenwoordigt de minimale hoeveelheid verandering gemeten bij een meetinstrument, waarvan we zeker zijn dat het geen artefact is van systematische error. In een grootschalige systematische review vonden we dat enkel 4.14% van de validiteitsartikelen rapporteren over meetfout, en dat de meetfout daarnaast berekent kon worden voor 3.82% van de artikelen. Om de implicaties van meetfout op klinisch onderzoek te illustreren, werd een simulatie studie uitgevoerd. Simulaties werden uitgevoerd op een hypothetische randomized controlled trial gericht op de behandeling van depressie zoals gemeten door de BDI-II. Beginwaarden en een vermindering van symptomen bij onbehandelde depressie (controle conditie) werden geëxtraheerd uit de literatuur. De Minimal Clinically Important Difference (MCID) werd gebruikt als maat van effectgrootte voor de verdere afname van symptomen over tijd bij behandelde depressie (behandel conditie). Drie parameters werden systematisch gevarieerd binnen de simulaties: steekproefgrootte (250 / 500 / 750), effectgrootte (0*MCID / 1*MCID / 2*MCID / 3*MCID) en meetfout (0% / 10% / 20% / 30% / 40%). Elke combinatie van parameters werd 5000 keer gesimuleerd. Relatieve bias is de afwijking van de coëfficiënt van belang. De relatieve bias werd meer afwijkend van bijna 0 (zonder meetfout) naar -0.5 (met een meetfout van 30% en 40%). Daarnaast lieten effectgroottes meer relatieve bias zien. ETA squared is een maat van effectgrootte. De ETA squared reikt van 0 tot 0.525 wanneer er 0% meetfout is, afhankelijk van de parameter van effectgrootte. Iedere

ETA squared gleed dichter naar nul voor meer toegevoegde meetfout. De resultaten van de simulatie toonden een stijging in bias met de toevoeging van meer meetfout. Daarnaast leek dit effect sterker voor grotere effectgroottes. Het resultaat van deze bias is een afname van effectgrootte, iets wat vooral ongunstig is bij 20% of meer meetfout. Het lijkt erop dat meetfout de mogelijkheid om een echt effect te detecteren beïnvloed.

De eHealth Impact Questionnaire (eHIQ) verschaft een gestandaardiseerde methode om de houding van eHealth gebruikers ten opzichte van eHealth te meten. Het is eerder gevalideerd in een populatie van eHealth gebruikers uit het Verenigd Koninkrijk en bestaat uit 2 delen en 5 subschalen. Deel 1 meet houdingen ten opzichte van eHealth in het algemeen en bestaat uit de subschalen: *Attitudes omtrent online gezondheidsinformatie* (5 items) en *Attitudes omtrent het online delen van gezondheidservaringen* (6 items). Deel 2 meet de houding ten opzichte van een specifieke eHealth applicatie en bestaat uit de subschalen *Vertrouwen en identificatie* (9 items), *Informatie en presentatie* (8 items) en *Begrip en motivatie* (9 items). De eHIQ is vertaald en gevalideerd in overeenstemming met de COSMIN criteria. De validatie bestond uit 3 steekproeven met in totaal 1287 deelnemers. Structurele validiteit werd vastgesteld door middel van confirmatieve factor analyses en exploratieve factor analyses. Interne consistentie werd beoordeeld met hierarchische omega (in alle 3 de steekproeven). Test-hertest betrouwbaarheid werd vastgesteld na 2 weken, waarbij gebruik gemaakt werd van tweewegs-intraclass correlatie coefficienten (steekproef 1). Meetfout werd beoordeeld door de kleinst waarneembare verandering te berekenen (steekproef 1). Convergente en divergente validiteit werden beoordeeld door middel van correlaties met overige variabelen (alle 3 de steekproeven). Een graded response model werd toegepast en item informatie curves werden weergegeven om de informatie per item over item trait levels te beschrijven (alle 3 de steekproeven). De originele factor structuur liet een slechte fit zien bij alle drie de steekproeven. EFAs lieten een goede fit zien voor een gemodificeerde factor structuur in de eerste steekproef. Deze factor structuur werd daarna getest in steekproef 2 en 3 en liet een aanvaardbare tot goede fit zien. Interne consistentie, test-hertest betrouwbaarheid, convergente validiteit en divergente validiteit waren aanvaardbaar tot goed voor zowel de originele als de gemodificeerde factor structuur, behalve voor test-hertest betrouwbaarheid van één van de originele subschalen en de twee afgeleide subschalen in de gemodificeerde factor structuur. De graded response model liet zien dat sommige items verminderd presteren in zowel de originele als de gemodificeerde factor structuur. De Nederlandse versie van de eHIQ (eHIQ-NL) laat een andere factor structuur zien in vergelijking met de Engelse versie. Deel 1 van de eHIQ-NL bestaat uit 3 subschalen: *Attitudes omtrent online gezondheidsinformatie* (5 items), *Comfort omtrent het delen van gezondheidservaringen* (3 items) en *Nut van het online delen van gezondheidservaringen* (3 items). Deel 2 van de eHIQ-NL bestaat uit drie subschalen: *Motivatie en vertrouwen om te handelen* (10 items), *Informatie en presentatie* (13 items) en *Identificatie* (3 items).

Kennis over symptoomclusters zou gerichte interventies kunnen informeren. We onderzochten symptoomclusters van kankerpatiënten door middel van machine learning technieken op een grote dataset. Hiervoor werd data gebruikt van kankerpatiënten die deelnamen aan de volledig geautomatiseerde online applicatie het Onkokompas. Deze applicatie was ondersteunend in hun zelfmanagement, door 1) hun symptomen te monitoren door middel van PROMs; 2) een gerichte terugkoppeling te geven op aan de hand van hun scores met een persoonlijk overzicht van ondersteunende zorgopties, gericht op het verminderen van symptomen en het verbeteren van gezondheid-gerelateerde kwaliteit van leven. In onze studie werd data over 26 algemene symptomen (fysiek en psychosociaal) meegenomen. Resultaten van de PROM van ieder symptoom worden aan de gebruiker gepresenteerd als geen welzijnsrisico, een gemiddeld welzijnsrisico, of een hoog welzijnsrisico. Data van 1032 kankerpatiënten werden geanalyseerd middels Hierarchical Densitiy-Based Spatial Clustering of Applications with Noise (HDBSCAN) op hoge risico scores en gemiddeld-tot-hoge risico scores afzonderlijk. Bij de analyse van de hoog risico scores werden zeven clusters onttrokken: één hoofdcluster met daarin de meest voorkomende fysieke en psychosociale symptomen en zes subclusters met verschillende combinaties van deze symptomen. Bij de analyse van de gemiddeld-tot-hoge risico scores werden drie clusters onttrokken: twee hoofdclusters werden geïdentificeerd, die onderscheid maakten tussen fysieke symptomen (en gevolgen daarvan) and psychosociale symptomen, en één subcluster met slechts problemen gerelateerd aan lichaamsgewicht. Er lijkt een inherent verschil te zijn in de co-morbiditeit van symptomen afhankelijk van de ernst van de symptomen. Bij kankerpatiënten met hoge risico scores liet de data een clustering met meer verbindingen tussen fysieke en psychosociale symptomen in verschillende subclusters zien. In kankerpatiënten met gemiddeld-tot-hoge risico scores zagen we minder verbindingen in de clustering van fysieke en psychosociale symptomen.

In hoofdstukken 2, 3 en 4 onderzochten we de psychometrische eigenschappen van drie PRMs. We beoordeelden de meerderheid van de psychometrische eigenschappen over deze drie PRMs ofwel als niet bepaalbaar (37.5%) ofwel inconsistent (25%); met iets meer dan één derde als voldoende (37.5%). We beoordeelden ook de kwaliteit van het bewijs voornamelijk als erg laag, laag of gemiddeld (81.8%), met de minderheid beoordeeld als hoog (18.2%). Bovendien, in een bredere systematische review naar de 29 PRMs gebruikt binnen het Oncokompas, vonden we dat voor een groot deel van deze PRMs onvoldoende informatie beschikbaar was met betrekking tot verschillende psychometrische eigenschappen. Dit is zorgwekkend, aangezien PRMs vaak toegepast worden in praktijk en wetenschap om informatie te verschaffen over de gezondheid van de patiënt en om gezondheidszorg te evalueren. In het bijzonder is er meer onderzoek noodzakelijk naar betrouwbaarheid, meetfout, responsiviteit en cross-culturele validiteit.

De validatie studie uitgevoerd betreffende de eHIQ (hoofdstuk 5) dient als een voorbeeld van hoe een validatie studie uitgevoerd kan worden volgens de COSMIN richtlijnen.

Hoofdstuk 6 illustreert de mogelijkheden om PRM data in te zetten voor het onderzoeken van relevante theoretische onderzoeksvragen. Door de toename van eHealth gebruik, en de inzet van PRMs door Nederlandse ziekenhuizen en zorgverzekeraars in de implementatie en focus op waarde-gedreven zorg, worden grote datasets met antwoorden op PRMs verzameld. Deze datasets kunnen ingezet worden om onderzoeksvragen te beantwoorden die normaal gesproken veel middelen vereisen om te onderzoeken. Het onderzoek naar symptoom clusters is een voorbeeld van een dergelijke onderzoeksvraag, en routinematig verzamelde data kan ingezet worden om aan deze onderzoekslijn bij te dragen. Routinematig verzamelde data kan tevens gebruikt worden voor validiteitsanalyses, met name in het onderzoek naar structurele validiteit, waarvoor bewijs vaak ontbreekt. Open datasets, gepubliceerd op platformen zoals Dataverse, LinkedScience en het Open Science Framework kunnen op eenzelfde manier ingezet worden. Onderzoek naar de test-hertest betrouwbaarheid, meetfout en responsiviteit vereisen een meer specifieke methodologische opzet. Om de inzet van middelen voor dergelijke studies te verminderen kan crowsourching worden ingezet. Resultaten van dergelijk onderzoek moet daarnaast toegankelijker overgebracht worden op de clinici en onderzoekers die daadwerkelijk gebruik maken van het meetinstrument. In een weerspiegeling van open data platforms en in lijn met de beweging van de Europese Unie naar open science, zou een platform kunnen worden ontwikkeld voor het uploaden van resultaten van validiteitsstudies, bij voorkeur inclusief de gebruikte dataset. Door gebruik te maken van machine readable formats zou dan een automatische, en kwalitatieve samenvatting van psychometrische eigenschappen kunnen worden gecreëerd.

Dankwoord

Dit proefschrift is alweer vier jaar in de maak. Terwijl ik hard gewerkt heb, heb ik dit niet kunnen doen zonder de hulp van veel mensen. Graag zou ik hen willen bedanken.

Allereerst wil ik Irma en Pim bedanken, mijn promotoren. Volgens mij had ik de jackpot van promotoren te pakken, want jullie waren ongelooflijk betrokken.

Irma, toen ik aan dit project begon zou er een postdoc aangenomen worden om mijn dag tot dag begeleiding op zich te nemen. Toen duidelijk werd dat de enige goede kandidaat toch niet beschikbaar was, heb ik me geen moment zorgen hoeven te maken. Jij was namelijk vanaf dag één altijd voor me beschikbaar, en ik hoefde nooit lang op een reactie te wachten. Als ik om een extra oog op mijn schrijfwerk vroeg - omdat dit mijn zwakke punt is - was dit nooit een probleem, en was je naast opbouwende feedback, altijd vol met complimenten. Wanneer ik twijfelde over de kwaliteit van mijn werk, wist je mij altijd met zachte hand gerust te stellen en samen te werken om het werk te verbeteren (zelfs wanneer jij vond dat het eigenlijk al best goed was).

Pim, jouw rasoptimisme was altijd aanstekelijk. Terwijl je kritisch was op het werk dat we uitvoerden, deed je dit altijd met een glimlach en de zekerheid dat alles goed zou komen. Ik kan me nog goed een moment van een aantal maanden geleden herinneren dat hier een perfect voorbeeld van is. We hadden een erg kritische peer reviewer gehad die het sterk oneens was met onze conclusie. Toen we het er over hadden of onze laatste revisie-ronde afdoende zou zijn, gaf ik aan dat “in het ergste geval schrijft deze reviewer een boos commentaar”. Met een stevige lach reageerde jij: “Mooi joh! Heb je meteen de eerste citatie!”.

I want to give a sincere thanks to all my colleagues at the section KNOP of the VU. I have been lucky to be able to work with you all. There was a culture where everyone was always available for help, comments, or just a chat. Cooperation was the name of the game. I hope I have been able to contribute to this atmosphere. It was a delight!

A special thanks goes to my roomies Felix, Mirjam, Niko, and in the last few months, Ángel. I can honestly say that without you, I would not have been able to finish this dissertation. You literally dragged me through the days when I was extracting data from my umpteenth article for my systematic review. Your support, and all the laughs we shared, is something I will cherish, and I already miss not seeing you on a weekly basis.

Het gezegde “it takes a village to raise a child” lijkt even toepasselijk op een systematische review. Ik wil graag (in no particular order) Femke, Anja, Heleen, Karen, Margot, Evalien, en Nienke bedanken voor hun hulp met screenen, data extraheren, en het

interpreteren van de data voor onze systematische review artikelen. Zonder jullie was dit proefschrift er niet geweest. Blijkbaar kost het een hele onderzoeksgroep om een systematische review te schrijven!

Er zijn een aantal docenten, vorige werkgevers en collega's tijdens mijn studie aan de Radboud Universiteit Nijmegen, waar ik graag wat woorden aan zou willen besteden. Het was een vreemde reis, als ik erover nadenk. Ik begon mijn bachelorstudie psychologie in Nijmegen, omdat ik geïnteresseerd was in wat mensen drijft. Ik had altijd een hekel gehad aan wiskunde op de middelbare school. Maar dankzij het fantastische curriculum van Jules Ellis en Inge Rabeling - van onderzoeksmethoden, tot statistiek, en tot psychometrie - besefte ik dat ik best goed was in het begrijpen, toepassen en zelfs uitleggen van deze concepten. De klas "Data-analyse" was mijn grootste eye-opener. Dit was het leukste vak dat ik tijdens mijn bachelor heb gevolgd. Aan de hand van mijn plezier in dit vak, heb ik gesolliciteerd als werkgroep-docent voor bijna ieder statistiek vak dat jullie gaven. Ik ben erg dankbaar dat jullie ervoor gekozen hebben om mij aan te nemen, omdat het uitleggen van statistiek en methodologie mij bepaalde concepten beter heeft laten begrijpen, en een belangrijke voorbereiding was op alles wat zou volgen.

Toen ik begon aan de Research Master: Behavioral Science, vertelde Inge me dat ze mensen tekort kwamen in de "Scriptiewerkplaats", een plek waar bachelor- en masterstudenten konden komen om statistisch advies in te winnen met betrekking tot hun scriptie. Inge heeft mijn naam geopperd in de groep, met enig scepticisme omdat ze "normaliter geen studenten aannemen voor deze positie". Lex Hendriks nodigde me uit op gesprek - waarschijnlijk omdat Inge me verder de hemel in had geprezen dan ik verdiende - en nam me kort daarna aan. Het was geweldig om omringd te zijn door collega's die tientallen jaren ervaring hadden met statistische analyses, en het is nog steeds gek hoe ik uiteindelijk mijn medestudenten (uit mijn eigen jaar) van statistisch advies heb mogen voorzien (en zelfs betaald kreeg om dit te doen)! Bedankt Inge, Pierre, Thea, Giovanni, William, Janneke, Mathieu en Kim, voor het verwelkomen van mij als een gelijke, ook al kon ik jullie ervaring en kennis niet evenaren. En natuurlijk speciaal veel dank aan Lex voor zijn vertrouwen dat ik studenten van goed advies kon voorzien. Mijn werk op de scriptiewerkplaats was één van de leukste en beste ervaringen van mijn tijd in Nijmegen.

Two paragraphs thanking people from Nijmegen, but I'm not done yet. During my time with the Research Master: Behavioural Science, I had the pleasure of having a couple of great lecturers and supervisors that were amazing. First of all, I want to thank Bernd Figner, Bill Burk, and Toon Cillesen for all the statistics classes, but especially for introducing me to R. I have become such an R fanboy over the past few years, that I might be able to teach you all some new tricks. I want to thank Bernd and Mike Rinck

for supervising my major and minor research projects, respectively. You taught me so many things, and even though I did not end up in experimental psychology for my PhD, these projects were incredibly formative. I still don't rule out the possibility of me returning to experimental psychology. I want to thank Ron Dotsch for taking the time to teach us how to use ggplot2, as well as how to design posters and presentations. These are skills that are often overlooked but have been very helpful during the past few years. Lastly, I want to thank all the other lecturers who provided the content for the theoretical courses, it was great to be introduced to so many fields of research.

Het is al erg lang geleden, maar ik zou ook graag mevrouw Meijer willen bedanken. Gedurende mijn tijd op de middelbare school waren er redelijk wat docenten die dachten dat ik niet veel zou bereiken. In mijn laatste twee jaar trad jij op als mijn mentrix, en jij leek je eigenlijk nooit zo druk om me te maken. De woorden "als je jouw best zou doen dan kom je nog wel ver, Koen" betekenden meer voor mij dan je misschien hebt gerealiseerd. Het heeft even geduurd voordat ik écht mijn best ging doen, maar het vertrouwen dat jij in me had heeft wat in gang gezet. Hopelijk heb je niet te veel te corrigeren aan dit Nederlandse dankwoord. Ik ben helaas nooit veel beter in onze moedertaal geworden.

Er zijn ook een aantal mensen die ik mijn excuses verschuldigd ben en wil bedanken voor hun geduld. Ik heb een aantal vrienden die de wereld voor me betekenen, en die ik de afgelopen paar jaar veel te weinig gezien heb doordat ik zo druk was. Dion, Hugo, en Thijmen, ik ga jullie vaker zien nu mijn promotie achter de rug is.

Ik ga nu heerlijk cliché zijn in het bedanken van mijn ouders. "Zonder mijn ouders had ik nooit kunnen promoveren" is zeker waar, want zonder mijn ouders had ik niet op dit blauwe bolletje rondgelopen, om artikelen te schrijven waar ik zeur over alles wat we beter zouden kunnen doen in de wetenschap. Slechte grapjes terzijde, bedankt pap en mam, jullie hebben altijd gezorgd voor een plek die ik thuis kan noemen. Jullie hebben mij altijd gesteund in mijn keuzes, zelfs toen deze nogal onverstandig waren. De meeste mensen die willen studeren in Nederland maken het VWO af, maar ik vond al die vakken maar saai, dus wou na het afmaken van de HAVO een andere route nemen. Ik ging naar het HBO om mijn propedeuse te halen in toegepaste psychologie om daarna door te stromen naar de universiteit. Nogal risicovol, want de propedeuse moest in één jaar gehaald worden, en daarna moest ik ook nog ingeloot worden. De grote grap, met die propedeuse kon ik alleen doorstromen naar een Bachelor in psychologie. Dus als ik niet ingeloot zou worden, of de studie niet leuk zou vinden, zou ik weinig mogelijkheden hebben. Ook al vonden jullie het zeker verstandiger als ik het VWO zou afmaken, steunden jullie mijn keuze en zorgden jullie voor de beste omgeving waar ik kon bereiken wat ik voor ogen had. Dit is natuurlijk maar één voorbeeld, maar dit

hele dankwoord zou veel te lang worden als ik jullie voor alles zou bedanken dat jullie voor mij hebben betekend. En ik kan er natuurlijk altijd op rekenen dat mam over me opschept, zodat ik kan doen alsof ik heel bescheiden ben.

Grote broer, soms vraag ik me af of je weet hoeveel invloed jij op mij hebt gehad in de afgelopen 29 jaar. Al qua muziekkeuze liep ik altijd enigszins achter jou aan tijdens onze tienerjaren. Ook bij veel van mijn hobby-keuzes (behalve misschien muziek) kan ik zien dat het hobby's zijn waar ik vaak aan ben begonnen omdat jij het leuk vond. Door jouw interesse in de wetenschap keken we veel meer naar Discovery Channel dan ik toen misschien had gewild. En je had altijd zoveel plezier in het uitleggen van de dingen die ik niet begreep. Ook al ben ik uiteindelijk een "zachte" wetenschap ingedoken, weet ik wel dat mijn interesse in wetenschap bij jou begonnen is. Bedankt Bart, omdat je zo'n fantastische grote broer bent.

Lieve Marije, na zoveel jaar is het grappig hoe erg we naar elkaar toe gegroeid zijn. Toen we 7 jaar bij elkaar waren besloten we allebei dezelfde Research Master te doen. Twee jaar later besloten we allebei tegelijkertijd te starten aan onze promotie. Wat hebben wij de afgelopen jaren toch vaak helemaal kapot samen op de bank gezeten. Op elkaar leunend als een paar invaliden hebben we elkaar er doorheen gesleept. Over de jaren heen ben ik het gezegde "zij is mijn rots in de branding" meer gaan waarderen. Na 13 jaar, heb jij me door zoveel moeilijke tijden gesleept, dat jij daadwerkelijk mijn rots in de branding bent, en eigenlijk altijd bent geweest. De momenten van rust en liefde, zijn mij de wereld waard geweest. Terwijl ik op het punt sta om dit proefschrift naar de drukker te sturen, ben jij druk bezig met jouw eigen artikelen. Ik ben zowaar trotser op jou, dan op dit proefschrift. Ik kan niet wachten op wat ons hierna te wachten staat. Ik weet dat we eindelijk weer meer tijd voor elkaar zullen hebben.

Als laatste, is er één iemand die mij altijd wist te laten ontspannen met ruw enthousiasme, onconditionele liefde, en iets te veel speeksel. Elsie, als we thuiskomen na mijn promotie, heb jij ook wat verdiend. Wat dacht je van een sappige hamburger of biefstuk?!

E

About the author

Written by Marije van der Hulst and Felix Bolinski

Koen Neijenhuijs (02-09-1990, Deventer, the Netherlands) did not burn the midnight oil - as many PhD students do - due to his lifelong experience with "working smarter, not harder". He developed this strategy as early as during kindergarten, where he would often take crafts home that were made by other children. Why waste time making these "stupid" crafts himself if the other parents would throw them away anyway?

In high school, at the Etty Hillesum Lyceum, Koen would follow this credo by strategically choosing to follow Higher General Secondary Education (Dutch: HAVO) instead of pre-university education (Dutch: VWO). This gave him ample time to dedicate to drumming, leaving open the possibility to go to a conservatory. Eventually deciding against that option, Koen followed a freshman year of Higher Professional Education (Dutch: HBO) and entered Radboud University Nijmegen, where he completed his Bachelor's degree (2013) in psychology with no time lost. Already during his undergraduate studies, Koen became a teaching assistant and developed a reputation as a statistical genius. One that was always happy to help and explain complex concepts to others. A reputation he would keep up for the entirety of his academic career. During his Research Master in Behavioral Science (2015, cum laude) at the Radboud University Nijmegen, Koen completed two additional Honors Academy degrees. He collaborated with other students from various disciplines on the topic of wider implications of neuroscience (2014) and travelled to the University of Glasgow (2015) in order to visit Dale Barr, PhD, fostering his enduring appreciation of the statistical program R.

During his PhD at the Vrije Universiteit Amsterdam, his strategical thinking and statistical knowledge were particularly valuable. For himself, but also for students, colleagues, and colleagues of colleagues. Rather than compiling weekly reports by hand, Koen invested a number of days in automating data extraction and reports at the start, a decision that would benefit him during the remainder of his PhD. His organizational skills enabled him not only to successfully finish his dissertation, but also in contributing to 10 papers as a co-author and serving as a data-analyst / consultant for War Child at

the same time. However, given Koen's helpful nature, his organization skills were merely a means to an end.

With his dissertation on "Patient reported measures in eHealth: on measurement properties and data opportunities", Koen shows that he does not shy away from complicated and difficult topics and that he is able to pinpoint relevant factors within a limited amount of time. He is a meta-science idealist, who believes that scientists are obliged to perform their research in accordance with the highest standards, or at least report reasons for deviations from these standards.

In 2020, Koen continued his career as data science consultant with the company Ordina.

Dear Koen, congratulations with the results of your hard work. We are very proud of you!

Marije van der Hulst & Felix Bolinski

E

Supplement

Chapter 2: The measurement properties of the IIEF

Supplementary Table 7.1. Internal consistency (Cronbach's Alpha) of the IIEF.

Reference	IIEF5	Total score	EF	OF	SD	IS	OS	Rating	Quality
IIEF-15									
Bayraktar et al. (2012) [66]			.92 - .93	.91 - .92	.90 - .92	.91 - .92	.91 - .92	Sufficient	Very good
Bayraktar et al. (2013) [67]		.78 - .92						Indeterminate	Inadequate*
Coyne et al. (2010) [72]			0.82	0.83	0.89	0.55	0.42	Insufficient	Very good
Dargis et al. (2013) [101]		.90 - .91	.88 - .89	.81 - .82	.84 - .87	.61 - .66	.86 - .87	Sufficient	Very good
Gonzáles et al. (2013) [76]		0.89	0.86	0.63	0.77	0.5	0.73	Insufficient	Adequate
Kriston et al. (2008) [78]			0.93	0.9	0.83	0.82	0.83	Sufficient	Very good
Lim et al. (2003) [102]		0.96	0.93	0.86	0.8	0.82	0.96	Insufficient	Very good
Nimbi et al. (2018) [81]					0.86		0.92	Sufficient	Adequate
Pascoal et al. (2017) [85]		0.85						Sufficient	Inadequate*
Quek et al. (2002) [86]			0.76	0.74	0.87	0.78	0.84	Sufficient	Inadequate**
Quinta Gomes et al. (2012) [87]			0.86	0.79	0.72	0.74	0.86	Sufficient	Very good
Rosen et al. (1997) [52]		.91 - .96	.92 - .96	.92 - .99	.77 - .91	.73 - .88	.74 - .87	Insufficient	Very good
Rubio-Aurioles et al. (2009) [89]				0.97				Insufficient	Very good
Tang et al. (2018) [92]		0.83						Sufficient	Inadequate*
Wiltink et al. (2003) [94]		0.95						Indeterminate	Adequate
IIEF-5									
Dargis et al. (2013) [101]	.83 - .85							Indeterminate	Very good
Lim et al. (2003) [102]	0.9							Indeterminate	Very good
Mahmood et al. (2012) [97]	0.88							Indeterminate	Inadequate**
Tang et al. (2015) [98]	0.64							Indeterminate	Very good
Utomo et al. (2015) [99]	0.94							Indeterminate	Adequate

Supplementary Tables

Supplementary Table 7.2. Measurement Error (Standard Error of Measurement & Smallest Detectable Change) of the IIEF.

Reference	IIEF.5	EF	OF	SD	IS	OS	Rating	Quality
IIEF-15								
Quek et al. (2002) [86]							Indeterminate / Insufficient (for EF only)	Inadequate*
SEM		3.59	1.34	0.69	1.88	2.96		
SDC		9.94	3.7	1.9	5.21	8.21		
Rosen et al. (1997) [52]							Indeterminate	Adequate
SEM		0.8 - 1.2	0.5 - 0.6	0.3	0.4 - 0.6	0.3 - 0.4		
IIEF-5								
Utomo et al. (2015) [99]							Indeterminate	Adequate
LoA	10.1							

Supplementary Table 7.3. Known-group validity of the IIEF.

Reference	Comparison groups	Outcome	Rating	Quality
IIEF-15				
Lim et al. (2003) [102]	ED patients vs non-patients	Patients with ED scored significantly lower on all scales of IIEF-15.	Sufficient	Adequate
Quek et al. (2002) [86]	Surgical treated group vs control group	Significant differences were found in the IIEF-15 domains of overall erectile function, overall sexual drive, overall intercourse satisfaction and overall sexual satisfaction score.	Sufficient	Inadequate*
Quinta Gomes et al. (2012) [87]	ED patients vs non-patients	Individuals from the clinical groups presented significantly lower scores on the five IIEF-15 domains when compared with controls	Sufficient	Adequate
Rosen et al. (1997) [52]	ED patients vs non-patients	Differences between domain scores of the IIEF-15 between the groups were greatest for the erectile function domain, followed by intercourse satisfaction and overall satisfaction. The least degree of difference between patients and controls was seen for the sexual desire domain, with results failing to reach statistical significance in study C.	Sufficient	Adequate
Tang et al. (2018) [92]	PE patients vs non-patients	PE patients scored significantly lower than controls the total score and subscale scores of the IIEF-15	Sufficient	Adequate
Tang et al. (2018) [92]	LPE patients vs APE patients	APE patients scored significantly lower than LPE patients on the total score, but not the subscale scores of the IIEF-15	Sufficient	Adequate
Wiltink et al. (2003) [94]	ED patients vs Peyronie patients vs non-patients	ED patients scored significantly lower than controls and Peyronie's patients on the IIEF-15 scales.	Sufficient	Adequate
IIEF-15 & IIEF-5			Sufficient	
Dargis et al. (2013) [101]	Age (65-74 years vs = 75)	At T1, significant differences on the IIEF-15 total and EF scales, but not on IIEF-5. At T2, significant differences on the IIEF-15 total, EF scales, and IIEF-5. In all cases older men scored lower.	Sufficient	Adequate
IIEF-5				
Rosen et al. (1999) [53]	ED patients vs non-patients	ED patients scored significantly lower than controls on the IIEF-5 total score	Sufficient	Doubtful**

DrTech = Doctor Technical Skills; DrInt = Doctor Interpersonal Skills; DrInfo = Doctor Information Provision; DrAva = Doctor Availability; NTech = Nurse Technical Skills; NInt = Nurse Interpersonal Skills; NInfo = Nurse Information Provision; NAva = Nurse Availability; SInt = Other Staff Interpersonal Skills; WT = Wait Times; HA = Hospital Access; ? = Indeterminate

Supplementary Table 7.4. Convergent validity of the IIEF.

Reference	Comparison	Correlations	Rating	Quality
IIEF-15				
Cappelleri et al. (2000) [70]	Self-assessment of ED & IIEF-EF at baseline (N=247)	0.65	Sufficient	Doubtful*
	Self-assessment of ED & IIEF-EF at week 12 (N=238)	0.86		
Cappelleri et al. (2009) [71]	EHS & IIEF erectile function	>=.75	Insufficient	Doubtful**
	EHS & IIEF intercourse satisfaction	>=.50		
	EHS & IIEF orgasmic function	>=.25		
	EHS & IIEF overall satisfaction	>=.25		
	EHS & IIEF sexual desire	.19-.49		
Flynn et al. (2013) [73]	PROMIS sexual interest & IIEF sexual desire	0.82	Sufficient	Adequate
	PROMIS erectile function & IIEF erectile function	0.83		
	PROMIS satisfaction & IIEF satisfaction	0.83		
	PROMIS orgasm & IIEF orgasmic function	0.62		
García-Cruz et al. (2011) [74]	EHS & IIEF erectile function	0.83	Sufficient	Doubtful***
Gelhorn et al. (2017) [75]	HIS-Q-SF & IIEF related subdomains	>.30	Sufficient	Doubtful****
	HIS-Q-SF sexual and libido & IIEF sexual desire	0.49		
Hwang et al. (2010) [77]	EHS & IIEF erectile function	0.79	Sufficient	Adequate
	EHS & IIEF intercourse satisfaction	0.62		
	EHS & IIEF overall satisfaction	0.61		
	Quality of Erection Questionnaire	0.7		
Maasoumi et al. (2017) [79]	SQOL-M total score & IIEF total score	0.5	Sufficient	Adequate
	SQOL-M total score & IIEF erectile function	0.37		
	SQOL-M total score & IIEF sexual desire	0.41		
	SQOL-M total score & IIEF intercourse satisfaction	0.44		
	SQOL-M total score & IIEF orgasmic function	0.43		
	SQOL-M total score & IIEF overall satisfaction	0.56		
Mulhall et al. (2008) [80]	SEQ erection & IIEF erectile function	0.7	Sufficient	Adequate
	SEQ individual satisfaction & IIEF intercourse satisfaction	0.51		
	SEQ individual satisfaction & IIEF overall satisfaction	0.7		
	SEQ couple satisfaction & IIEF intercourse satisfaction	0.45		
	SEQ couple satisfaction & IIEF overall satisfaction	0.67		

Reference	Comparison	Correlations	Rating	Quality
IIEF-15				
Nimbi et al. (2018) [81]	SMQ automatic thoughts & IIEF erectile function	-.10 - -.34	Insufficient	Adequate
	SMQ automatic thoughts & IIEF orgasmic function	-.07 - -.26		
	SMQ automatic thoughts & IIEF sexual desire	-.10 - -.32		
	SMQ automatic thoughts & IIEF intercourse satisfaction	-.12 - -.34		
	SMQ automatic thoughts & IIEF overall satisfaction	-.19 - -.45		
	SMQ automatic thoughts & IIEF total score	-.13 - -.38		
O'Toole et al. (2018) [83]	IBD-MSDS & IIEF erectile function	-0.27	Insufficient	Adequate
	IBD-MSDS & IIEF orgasmic function	-0.26		
	IBD-MSDS & IIEF sexual desire	-0.31		
	IBD-MSDS & IIEF intercourse satisfaction	-0.18		
	IBD-MSDS & IIEF overall satisfaction	-0.27		
Parisot et al. (2014) [84]	EHS & IIEF erectile function	.66 - .68	Sufficient	Doubtful*****
	EHS & IIEF orgasmic function	.45 - .49		
	EHS & IIEF sexual desire	.29 - .50		
	EHS & IIEF intercourse satisfaction	.58 - .59		
	EHS & IIEF overall satisfaction	.60 - .66		
Pascoal et al. (2017) [85]	BASEF & IIEF total score	-0.24	Insufficient	Adequate
Rosen et al. (1997) [52]	Clinician rating ED & IIEF erectile function	0.75	Sufficient	Adequate
	Clinician rating ED & IIEF orgasmic function	0.51		
	Clinician rating ED & IIEF sexual desire	0.61		
	Clinician rating ED & IIEF intercourse satisfaction	0.45		
	Clinician rating ED & IIEF overall satisfaction	0.63		
Rubio-Aurioles et al. (2009) [89]	Clinical diagnosis ED	0.813	Sufficient	Adequate
	Female Assessment of Male Erection	0.78		
Saffari et al. (2016) [90]	MGSIS & IIEF erectile function	0.31	Insufficient	Adequate
	MGSIS & IIEF orgasmic function	0.28		
	MGSIS & IIEF sexual desire	0.4		
	MGSIS & IIEF intercourse satisfaction	0.34		
	MGSIS & IIEF overall satisfaction	0.39		
Tang et al. (2018) [92]	PEDT & IIEF total score	-.23 - -.38	Insufficient	Adequate

Reference	Comparison	Correlations	Rating	Quality
IIEF-15				
	PEDT & IIEF erectile function	-.18 - -.32		
	PEDT & IIEF intercourse satisfaction	-.17 - -.29		
	PEDT & IIEF orgasmic function	-.01 - -.08		
	PEDT & IIEF sexual desire	.06 - .15		
	PEDT & IIEF overall satisfaction	-.33 - -.58		
Wiltink et al. (2003) [94]	Partner satisfaction & IIEF total score	0.31	Sufficient	Adequate
	Partner satisfaction & IIEF sexual function	0.27		
	Clinical rating ED & IIEF total score	0.684		
	Clinical rating ED & IIEF sexual function	.69.		
IIEF-5				
Aslan et al. (2011) [95]	Erection Hardness Scale	0.38	Insufficient	Adequate
Cappelleri et al. (2001) [100]	Self-assessment of ED at Baseline (N=247)	0.66	Sufficient	Doubtful*
	Self-assessment of ED at 12 weeks (N=238)	0.86		
	EDITS (N=237)	0.68		
	EDITS partner (N=78)	0.56		
	Global efficacy of erections (N=236)	0.69		

Supplementary Table 7.5. Divergent validity of the IIEF.

Reference	Comparison instrument	Correlations	Rating	Quality
IIEF-15				
Rosen et al. (1997) [52]	Locke-Wallace Marital Adjustment Test; social desirability	None of the correlations between domain scores and measures of marital adjustment or social desirability reached statistical significance	Sufficient	Doubtful*
Wiltink et al. (2003) [94]	STAI; CES-D; Body Complaints; SDS-CM	None of the correlations between the total scale of erectile function scale and comparison instruments reached statistical significance	Sufficient	Adequate
IIEF-15 & IIEF-5				
Dargis et al. (2013) [101]	Dyadic Adjustment Scale, SF-12	.06 - .33	Sufficient	Adequate

DrTech = Doctor Technical Skills; DrInt = Doctor Interpersonal Skills; DrInfo = Doctor Information Provision; DrAva = Doctor Availability; NTech = Nurse Technical Skills; NInt = Nurse Interpersonal Skills; NInfo = Nurse Information Provision; NAva = Nurse Availability; SInt = Other Staff Interpersonal Skills; WT = Wait Times; HA = Hospital Access; IE = Information Exchange; HC = Hospital Comfort; OA = Overall Satisfaction; + = Sufficient

Supplementary Table 7.6. Responsiveness of the IIEF.

Reference	Time period	Results	Rating	Quality
IIEF-15				
Althof et al. (2006) [65]	12 weeks	Sildenafil demonstrated significantly (p < .001) greater mean improvements (95% CI) over placebo on IIEF-EF: 11.7 (10.4-13.0) vs 5.2 (3.9-6.5)	Sufficient	Adequate
Cappelleri et al. (2000) [70]	12 weeks	For each measure, sildenafil resulted in a singificant benefit of placebo between pre to post (p < .0001)	Sufficient	Adequate
O'Leary et al. (2006) [82]	12 weeks	Statistically significant differences between the sildenafil and placebo group were observed for mean change from pre to post of the EF domain (9.3 [95% CI 7.9-10.7] vs 3.6 [95% CI 2.2-5.0]), all other IIEF domains, and all inidividual IIEF questions except frequency of attempted intercourse	Sufficient	Adequate
Parisot et al. (2014) [84]	6 – 12 months	The following scores showed to be significantly improving over the study period: IIEF-EF without and with treatment, and IIEF-OS (effect sizes: +0.4 [P = 0.006], +0.3 [P = 0.008], +0.2 [P = 0.03], respectively).	Sufficient	Adequate
Quek et al. (2002) [86]	3 months	Significant mean changes were observed in the item of ejaculation frequency (Cohen's d = .39) and overall domain of orgasmic function (Cohen's d = .30). The lowest magnitude of change was noted in domain of erectile function and sexual drive and item of intercourse frequency. In contrast, none of the comparisons in the control subjects approached significance.	Sufficient	Inadequate*
Rosen et al. (1997) [1]	12 weeks	Significant changes were observed across all five domains in the treatment responder group. The lowest magnitude of change was noted for the sexual desire domain. In contrast, none of the comparisons in the treatment nonresponder group approached significance.	Sufficient	Adequate
IIEF-5				
Cappelleri et al. (2001) [100]	12 weeks	The difference in mean change between treatment groups was significant (p < .001) with an effect size of 0.96.	Sufficient	Adequate
Utomo et al. (2015) [99]	6 months	The change in IIEF-5 score in treated patients after 6 months was 2.2 +- 3.9 compared to -0.6 +- 2.8 in untreated patients (p = 0.007).	Sufficient	Doubtful

Supplementary Table 7.7. Internal consistency (Cronbach's Alpha) of the FSFI.

Reference	Total score	DE	AR	LU	OR	SA	PA	Rating	Quality
FSFI-19									
Achimas-Cadariu et al. (2013) [109]		.80 - .89	.89 - .96	.96-.96	.93-.94	.97-.97	.89-.96	Sufficient	Very good
Anis et al. (2011) [111]		0.89	0.92	0.91	0.85	0.9	0.94	Sufficient	Very good
Bartula et al. (2015) [114]		.92 - .93	.93 - .94	.95 - .96	.92 - .94	0.89	.91 - .92	Sufficient	Very good
Baser et al. (2012) [115]	0.94	0.9	0.92	0.94	0.9	0.85	0.93	Sufficient	Very good
Burri et al. (2010) [184]	0.93	0.76	0.91	0.92	0.89	0.87	0.88	Sufficient	Very good
Carvalho et al. (2012) [185]	0.93	0.81	0.87	0.8	0.79	0.88	0.89	Sufficient	Very good
Chang et al. (2009) [122]	0.96							Insufficient	Inadequate
Clayton et al. (2010) [124]	0.95	.46-.52						Insufficient	Doubtful
Fakhri et al. (2012) [128]	0.86	0.81	0.76	0.82	0.9	0.87	0.72	Sufficient	Very good
Filocamo et al. (2014) [131]	0.97	0.92	0.95	0.96	0.94	0.92	0.92	Insufficient	Very good
Forbes et al. (2014) [104]		0.88		0.89	0.93	0.89	0.89	Sufficient	Very good
Gerstenberger et al. (2010) [133]		0.92						Sufficient	Very good
Ghassamia et al. (2013) [132]		0.82				0.89	0.95	Sufficient	Doubtful
Hevesi et al. (2017) [137]	.89 - .93	.79 - .92	.83 - .92	.80 - .93	.89 - .85	.80 - .86	.94 - .91	Sufficient	Very good
Kalmbach et al. (2015) [140]		0.89	0.87	0.81	0.88	0.87	0.86	Sufficient	Very good
Likes et al. (2006) [141]		.86 - .91	.94 - .97	.96 - .98	.97 - .97	.81 - .85	.97 - .98	Insufficient	Adequate
Liu et al. (2016) [143]	0.94	0.72	0.89	0.9	0.83	0.85	0.89	Sufficient	Very good
Meston et al. (2003) [145]	.89 - .92	.58 - .84	.83 - .91	.85 - .95	.89 - .90	.74 - .84	.90 - .94	Sufficient	Very good
Nowosielski et al. (2013) [150]	.87 - .96	.82 - .88	.87 - .92	.86 - .92	.82 - .88	.71 - .91	.90 - .96	Sufficient	Very good
Opperman et al. (2013) [151]	0.81	0.84	0.8	0.82	0.95	0.78	0.84	Sufficient	Adequate
Rehman et al. (2015) [154]	0.95	0.96	0.97	0.89	0.84	0.96	0.97	Insufficient	Very good
Rillon-Tabil et al. (2013) [156]	0.95	0.88	0.92	0.88	0.79	0.91	0.76	Sufficient	Adequate
Rosen et al. (2000) [54]	0.97	0.92	0.95	0.96	0.94	0.89	0.94	Sufficient	Very good
Ryding et al. (2015) [159]	.81 - .96	.77 - .93	.67 - .93	.80 - .96	.78 - .90	.53 - .89	.67 - .96	Sufficient	Very good
Sidi et al. (2007) [161]	.87 - .97	.54 - .87	.76 - .93	.78 - .93	.72 - .90	.85 - .95	.84 - .94	Sufficient	Very good
Stephenson et al. (2016) [163]		0.93	0.93	0.93	0.9		0.86	Sufficient	Adequate
Sun et al. (2011) [164]	.84 - .91	.71 - .89	.74 - .92	.90 - .93	.69 - .89	.84 - .92	.77 - .94	Sufficient	Very good

Reference	Total score	DE	AR	LU	OR	SA	PA	Rating	Quality
FSFI-19									
Takahashi et al. (2011) [165]	0.97	0.92	0.96	0.97	0.95	0.84	0.96	Insufficient	Very good
Ter Kuile et al. (2006) [166]	.93 - .98	.72 - .90	.88 - .96	.96 - .97	.83 - .95	.80 - .87	.84 - .98	Sufficient	Very good
Trudel et al. (2012) [167]	.90 - .91	.85 - .86	.86 - .90	.82 - .86	.83 - .85	0.73	.80 - .82	Sufficient	Very good
Vallejo-Medina et al. (2018) [169]		0.84	0.84	0.85	0.86	0.85	0.89	Sufficient	Very good
Verit et al. (2007) [171]	.95 - .97	.91 - .93	.90 - .91	0.91	.89 - .91	.91 - .93	.92 - .95	Sufficient	Very good
Wiegel et al. (2005) [173]	.93 - .95	.88 - .91	.91 - .96	.94 - .97	.91 - .93	.82 - .89	.95 - .98	Sufficient	Very good
Witting et al. (2008) [174]	0.95	.72 - .73	0.92	0.96	.90 - .91	0.88	0.96	Sufficient	Very good
Wylomanski et al. (2014) [176]	0.97	0.88	0.96	0.97	0.94	0.84	0.97	Insufficient	Very good
Zachariou et al. (2017) [177]	0.92							Sufficient	Inadequate
FSFI-BC (34 items)									
Bartula et al. (2015b) [183]		.81 - .94		.87 - .97	.92 - .98	.71 - .90	.89 - .96	Sufficient	Very good
FSFI-LL									
Burri et al. (2010) [184]	0.92	0.79	0.89	0.83	0.87	0.89	0.83	Indeterminate	Very good
FSFI-6									
Chedraui et al. (2012) [179]	0.91							Indeterminate	Very good
Isidori et al. (2010) [55]	0.79							Indeterminate	Very good
Lee et al. (2014) [180]	0.89							Indeterminate	Inadequate
Perez-lopez et al. (2012) [182]	0.91							Indeterminate	Very good

FSFI: Female Sexual Function Index; DE: Desire; AR: Arousal; LU: Lubrication; OR: Orgasm; SA: Satisfaction; PA: Pain; BC: Breast Cancer

Supplementary Table 7.8. Known-group validity of the FSFI.

Reference	Comparison groups	Outcome	Rating	Quality
FSFI-19				
Achimas-Cadariu et al. (2013) [109]	Cervical conization patients vs controls	Significant differences between women with cervical conization and controls on all subscales, except orgasm	Sufficient	Adequate
Anis et al. (2011) [111]	FSD patients vs controls	Statistically significant differences were found between mean scores (ArFSFI total score and scores of all domains) of the case group (women with sexual dysfunction) and those from the noncase group (women without sexual dysfunction)	Sufficient	Adequate
Baser et al. (2012) [115]	Cancer treatment groups	Cancer survivors who received neither chemotherapy nor radiation had better scores on the FSFI-19 subscales Lubrication (p = .004), Pain (p = .046), and total scores (p = .058) compared to survivors who received chemotherapy, radiotherapy, or both.	Sufficient	Adequate
Clayton et al. (2010) [124]	HSDD patients vs controls	Women with HSDD scores lower than those without FSD. However, FSFI did not discriminate between women with HSDD and those with FSAD	Sufficient	Adequate
Fakhri et al. (2012) [128]	FSD patients vs controls	Women with FSD reported significantly lower FSFI scores in comparison with those without FSD (P < 0.001) (Table 3). After controlling for multiple comparisons, all of the FSD subscales remained statistically significant, with the exception of arousal.	Sufficient	Adequate
Ghassamia et al. (2013) [132]	Urology clinic patients vs controls	As expected, the healthy participants reported better sexual functioning than the clinic (patients) sample	Sufficient	Adequate
Likes et al. (2006) [141]	VIN patients vs controls	Significant differences on all subscales except pain between women with VIN and healthy controls	Sufficient	Adequate
Meston et al. (2003) [145]	FOD / HSDD patients vs controls	Significant differences between women with FOD and controls and between women with HSDD and controls on each of the FSFI domain and total scores	Sufficient	Adequate
Meston et al. (2005) [146]	FSD patients vs controls	FSFI domain scores differed significantly between FSD and control women	Sufficient	Adequate
Nowosielski et al. (2013) [150]	FSD patients vs controls	Significant differences in mean scores for all domains as well as in the total scores between women with FSD and those without FSD	Sufficient	Adequate
Rellini et al. (2006) [155]	FSAD patients vs HSDD / FOD patients	Results from these t-tests showed a significant difference in the arousal (t(120) = 2.55, P < 0.01), lubrication (t(120) = 3.05, P < 0.01), and satisfaction (t(120) = 3.18, P < 0.01) domains between women with FSAD and women with HSDD or FOD	Sufficient	Adequate
Rillon-Tabil et al. (2013) [156]	Diabetic patients vs controls	Significant difference between diabetic and non-diabetic women was noted with total FSFI score and the following domains: desire, arousal and satisfaction	Sufficient	Adequate

Reference	Comparison groups	Outcome	Rating	Quality
FSFI-19				
Rosen et al. (2000) [54]	FSD patients vs controls	Females with sexual arousal disorder and control participants had significant ($p < .05$) differences on all FSFI-19 subscales and total score. The largest differences between the groups were seen for the subscales Lubrication and Arousal.	Sufficient	Adequate
Ryding et al. (2015) [159]	HSDD patients vs controls	Statistically significant differences between women with HSDD and women without were observed for the total scale and for all of the domains	Sufficient	Adequate
Sidi et al. (2007) [161]	FSD patients vs controls	Significant difference ($P < 0.01$) between the mean scores (total and of each domain) from the case group (women with sexual dysfunction) and those from the noncase group (women without sexual dysfunction)	Sufficient	Adequate
Sun et al. (2011) [164]	FSD patients vs controls	Significant difference between the women with FSD and those without on each domain score aswell as the total score	Sufficient	Adequate
Takahashi et al. (2011) [165]	Premenopausal women vs postmenopausal women	The regular menstruation group showed significantly higher scores than the menopause group in the total score and the subdomains of desire, arousal, lubrication, orgasm, and pain (Table 5). However, the two groups did not differ significantly in the satisfaction domain	Sufficient	Adequate
Ter Kuile et al. (2006) [166]	FSD patients vs controls	All the six FSFI subscales scores and the FSFI-total score were significantly lower in the sample of women with a sexual complaint (FSD group) than in the sample of women without sexual complaints	Sufficient	Adequate
Trudel et al. (2012) [167]	Women >75 years old vs women 65-74 years old	Women aged 75 and over had lower scores than those aged between 65 and 74 years	Sufficient	Adequate
Verit et al. (2007) [171]	CPP patients vs controls	Significant differences ($p < .0001$) on all FSFI subscales between women with CPP and women without CPP	Sufficient	Adequate
Wiegel et al. (2005) [173]	FSD patients vs controls	Significant difference ($p<.001$) on all FSFI subscales between women with FSD and women without FSD	Sufficient	Adequate
Wylomanski et al. (2014) [176]	Premenopausal women vs postmenopausal women; married vs non-married women	Significant differences in summary score between premenopausal and postmenopausal women, and according to the marital status were found	Sufficient	Adequate
Zachariou et al. (2017) [177]	Women experiencing subjective sexual distress vs women not experiencing subjective sexual distress	Significant difference ($p<.01$) between women with and without subjective sexual distress associated with a sexual dysfunction	Sufficient	Adequate

IIEF: International Index of Erectile Function; EF: Erectile Function; OF: Orgasmic Function; SD: Sexual Desire; IS: Intercourse Satisfaction; OS: Overall Satisfaction * Due to an extremely small N

Supplementary Table 7.9. Convergent validity of the FSFI.

Reference	Comparison	Correlations	Rating	Quality
FSFI-19				
Ahmed et al. (2017) [110]	FSDS & FSFI desire	-0.6	Sufficient	Adequate
	FSDS & FSFI arousal	-0.4		
	FSDS & FSFI lubrication	-0.2		
	FSDS & FSFI orgasm	-0.7		
	FSDS & FSFI pain	-0.6		
	FSDS & FSFI satisfaction	-0.8		
	FSDS & FSFI total	-0.7		
Aydin et al. (2016) [112]	FSDS-R & FSFI total	-0.19	Insufficient	Adequate
	FSDS-R & FSFI desire	0.02		
	FSDS-R & FSFI arousal	-0.21		
	FSDS-R & FSFI lubrication	-0.09		
	FSDS-R & FSFI orgasm	-0.41		
	FSDS-R & FSFI satisfaction	-0.32		
	FSDS-R & FSFI pain	-0.13		
Azimi Nekoo et al. (2014) [113]	FSDS-R & FSFI dimensions	-.40 - .16	Indeterminate	Adequate
Bartula et al. (2015) [114]	CAncer Rehabilitation Evaluation System	-.48 - -.70	Sufficient	Adequate
	World Health Organization Quality Of Life - 100	.44 - .73		
Bartula et al. (2015b) [183]	CARES & FSFI-BC changes due to cancer	-.25 - -.38	Insufficient	Doubtful
	CARES & FSFI-BC desire	-.29 - -.47		
	CARES & FSFI-BC lubrication	-.24 - -.35		
	CARES & FSFI-BC orgasm	-0.34		
	CARES & FSFI-BC pain	-0.35		
	CARES & FSFI-BC satisfaction	-.39 - -.62		
	CARES & FSFI-BC distress	-.32 - -.36		
	SPS & FSFI-BC changes due to cancer	-.47 - -.58		
	SPS & FSFI-BC desire	-.49 - -.67		
	SPS & FSFI-BC lubrication	-.45 - -.47		
	SPS & FSFI-BC orgasm	-.45 - -.61		
	SPS & FSFI-BC pain	-.37 - -.42		

Reference	Comparison	Correlations	Rating	Quality
FSFI-19				
	SPS & FSFI-BC satisfaction	-.52 - -.56		
	SPS & FSFI-BC distress	-.60 - -.76		
Baser et al. (2012) [115]	Center for Epidemiologic Studies Depression	-.17 - -.40	Insufficient	Adequate
	Impact of Event Scale	-.04 - -.23		
	Menopausal Symptom Checklist	-.30 - -.51		
	Reproductive Concerns Scale	-.08 - -.30		
	Functional Assessment of Cancer Therapy - General	.15 - .30		
	Functional Assessment of Cancer Therapy - Cervix	.19 - .47		
	Medical Outcomes Short Form	.10 - .35		
	Adapted Dyadic Adjustment Scale	-.05 - .26		
Bloemendaal et al. (2015) [116]	FSFI Arousal & SESII-W Sexual Excitation	0.46	Insufficient	Adequate
	FSFI Arousal & SESII-W & arousability	0.48		
	FSFI Arousal & SESII-W & partner characteristics	0.29		
	FSFI Arousal & SESII-W & sexual power dynamics	0.37		
	FSFI Arousal & SESII-W & smell	0.24		
	FSFI Arousal & SESII-W & setting	0.3		
	FSFI Arousal & SESII-W & sexual inhibition	-0.46		
	FSFI Arousal & SESII-W & concerns about sexual function	-0.3		
	FSFI Arousal & SESII-W & arousal contingency	-0.57		
	FSFI Arousal & SESII-W & relationship importance	-0.16		
Bornefeld-Ettmann et al. (2018) [118]	SSEI & FSFI total	0.58	Sufficient	Adequate
	SSEI & FSFI desire	0.33		
	SSEI & FSFI arousal	0.48		
	SSEI & FSFI lubrication	0.41		
	SSEI & FSFI orgasm	0.4		
	SSEI & FSFI satisfaction	0.54		
	SSEI & FSFI pain	0.38		
Burri et al. (2010) [184]	FSFI desire & FSFI-LL desire	0.76	Sufficient	Adequate
	FSFI arousal & FSFI-LL arousal	0.49		

Reference	Comparison	Correlations	Rating	Quality
FSFI-19				
	FSFI lubrication & FSFI-LL lubrication	0.44		
	FSFI orgasm & FSFI-LL orgasm	0.51		
	FSFI satisfaction & FSFI-LL satisfaction	0.41		
	FSFI pain & FSFI-LL pain	0.5		
	FSFI total & FSFI-LL total	0.43		
Burri et al. (2018) [119]	SCS desire & FSFI desire	0.44	Sufficient	Adequate
	SCS arousal & FSFI arousal	0.41		
	SCS lubrication & FSFI lubrication	0.39		
	SCS orgasm & FSFI orgasm	0.57		
	SCS satisfaction & FSFI satisfaction	0.6		
	SCS vaginismus & FSFI pain	0.43		
	SCS dyspareunia & FSFI pain	0.69		
	SCS total & FSFI total	0.67		
Carpenter et al. (2015) [120]	FSDS-R & FSFI	FSDS-R items were modestly correlated with FSFI domains of desire and satisfaction, and poorly correlated with other FSFI domains.	Insufficient	Adequate
Clayton et al. (2006) [123]	SIDI-F & FSFI total	0.84	Sufficient	Adequate
	SIDI-F & FSFI arousal	0.82		
	SIDI-F & FSFI desire	0.86		
	SIDI-F & FSFI lubrication	0.51		
	SIDI-F & FSFI orgasm	0.37		
	SIDI-F & FSFI pain	0.34		
	SIDI-F & FSFI satisfaction	0.8		
Clayton et al. (2010) [124]	SIDI-F & FSFI Total	.62 - .77	Sufficient	Adequate
	SIDI-F & FSFI arousal	.56 - .75		
	SIDI-F & FSFI desire	.50 - .73		
	SIDI-F & FSFI lubrication	.28 - .59		
	SIDI-F & FSFI orgasm	.48 - .66		
	SIDI-F & FSFI pain	.16 - .44		

Reference	Comparison	Correlations	Rating	Quality
FSFI-19				
	SIDI-F & FSFI satisfaction	.54 - .69		
Constantine et al. (2017) [125]	PISQ-IR arousal/orgasm & FSFI desire	0.36	Sufficient	Adequate
	PISQ-IR arousal/orgasm & FSFI arousal	0.49		
	PISQ-IR arousal/orgasm & FSFI lubrication	0.5		
	PISQ-IR arousal/orgasm & FSFI orgasm	0.3		
	PISQ-IR arousal/orgasm & FSFI & pain	0.5		
	PISQ-IR partner related & FSFI satisfaction	0.48		
	PISQ-IR global quality rating & FSFI desire	0.3		
	PISQ-IR global quality rating & FSFI arousal	0.27		
	PISQ-IR global quality rating & FSFI satisfaction	0.41		
	PISQ-IR condition impact & FSFI arousal	0.4		
	PISQ-IR desire & FSFI desire	0.75		
DeRogatis et al. (2010) [126]	Womenâ			

IIEF: International Index of Erectile Function; ED: Erectile Dysfunction; EF: Erectile Function; PE = Premature Ejaculation; LPE = Lifelong Premature Ejaculation; APE = Acquired Premature Ejaculation * Due to an extremely small N ** Due to very unequal group sizes

Supplementary Table 7.10. Divergent validity of the FSFI.

Reference	Comparison.instrument	Correlations	Rating	Quality
FSFI-19				
Achimas-Cadariu et al. (2013) [109]	EORTC QLQ-C30	All correlations <.24	Sufficient	Adequate
Bartula et al. (2015) [114]	FSFI total & Body Image Scale	-.11 - -.33	Sufficient	Adequate
	FSFI total & Fatigue Assessment Scale	-.16 - -.25		
	FSFI total & Impact of Events Scale	-.11 - -.15		
	FSFI total & Depression Anxiety Stress Scales	-.11 - -.22		
	FSFI total & Medical Outcomes 20-item	.12 - .17		
	FSFI total & Revised Dyadic Adjustment Scale	.12 - .42		
Bartula et al. (2015b) [183]	Fatigue Assessment Scale; Body Image Scale; Medical Outcomes Study Physical Health Subscale; Medical Outcomes Study Mental Health Subscale; Revised Dyadic Adjustment Scale	With the exception of correlations between relationship adjustment and FSFI-BC satisfaction (r = 0.46), body image and FSFI-BC distress (r = -0.36), mental health and FSFI-BC satisfaction (r = 0.32), and body image and FSFI-BC changes after cancer (r = -0.31), all other correlations between the FSFIBC subscales and measures of body image, fatigue, physical health, mental health and relationship adjustment were below 0.30	Insufficient	Adequate
Likes et al. (2006) [141]	EORTC QLQ-C30 total score & FSFI desire	.20 - .21	Insufficient	Adequate
	EORTC QLQ-C30 total score & FSFI arousal	.05 - .26		
	EORTC QLQ-C30 total score & FSFI lubrication	.11 - .25		
	EORTC QLQ-C30 total score & FSFI orgasm	.02 - .32		
	EORTC QLQ-C30 total score & FSFI satisfaction	.05 - .36		
	EORTC QLQ-C30 total score & FSFI pain	.08 - .31		
	EORTC QLQ-C30 total score & FSFI total	.09 - .34		
Nowosielski et al. (2013) [150]	Likert & FSFI desire	0.39	Insufficient	Doubtful
	Likert & FSFI arousal	0.41		
	Likert & FSFI lubrication	0.32		
	Likert & FSFI orgasm	0.37		

Reference	Comparison.instrument	Correlations	Rating	Quality
FSFI-19				
	Likert & FSFI satisfaction	0.51		
	Likert & FSFI pain	0.27		
	Likert & FSFI total	0.45		
Rosen et al. (2000) [54]	LWMAT & FSFI Desire	.04 - .19	Sufficient	Doubtful
	LWMAT & FSFI Arousal	.19 - .43		
	LWMAT & FSFI Lubrication	.09 - .42		
	LWMAT & FSFI Orgasm	.03 - .37		
	LWMAT & FSFI Satisfaction	.40 - .72		
	LWMAT & FSFI Pain	.20 - .33		
	LWMAT & FSFI Total	.22 - .53		
Ryding et al. (2015) [159]	Symptom Checklist-90-Revised	All correlations <.20 except between the arousal domain and SCL-90-R GSI (r = -0.27), depression (r = -0.27), and anxiety (r = -0.29), between the lubrication domain and anxiety (r = -0.27), and between the arousal domain and SCL-90-R somatization (r = 0.32)	Sufficient	Doubtful
Trudel et al. (2012) [167]	Dyadic Adjustment Scale; Short Form Health Survey	DAS: All below .30 except for FSFI total and satisfaction. SF-12 all below .30	Sufficient	Doubtful

IIEF: International Index of Erectile Function; ED: Erectile Dysfunction; EHS: Erection Hardness Score; PROMIS: Patient-Reported Outcomes Measurement Information System; HIS-Q-SF: Hypogonadism Impact of Symptoms Questionnaire Short Form; SQOL-M: Sexual Quality of Life–Male; SEQ: Sexual Experience Questionnaire; SMQ: Sexual Modes Questionnaire; IBD-MSDS: Inflammatory Bowel Disease Male Sexual Dysfunction Scale; BASEF: Beliefs About Sexual Functioning Scale; MGSIS: Male Genital Self-Image Scale; PEDT: Premature Ejaculation Tool * Due to lack of information on measurement properties of comparator instrument ** Due to imprecise reporting of results *** Due to use of Pearson correlation where Spearman correlation was appropriate **** Due to imprecise reporting of hypotheses ***** Due to a small N

Chapter 4: The measurement properties of the EORTC IN-PATSAT32

Supplementary Table 7.11. Characteristics of included studies.

Reference	Population	Sample size	Main aim of study
Aboshaiqah et al., 2016 [208]	Saudi Arabian palliative cancer patients	130	Determine the relationship between quality of life and satisfaction with care among palliative cancer patients in Saudi Arabia
Arraras et al., 2009 [200]	Spanish cancer patients	80	Translate and validate the EORTC IN-PATSAT32 for the Spanish general cancer population
Arraras et al., 2010 [206]	International cancer patients	509	Test the structure, validity, and reliability of the EORTC QLQ-INFO25
Asadi-lari et al., 2015 [205]<U+2060>	Iranian cancer patients receiving radio- or chemotherapy	173	Translate and validate the EORTC QLQ-INFO25 for the Iranian general cancer population
Hjörleifsdóttir et al., 2010 [198]	Icelandic cancer patients receiving radio- or chemotherapy in an outpatient setting	217	Translate and validate the EORTC IN-PATSAT32 for the Icelandic general cancer population
Obtel et al, 2017 [203]	Moroccan hospitalized cancer patients	133	Translate and validate the EORTC IN-PATSAT32 for the Moroccan general cancer population
Pishkuhi et al., 2014 [199]	Iranian cancer patients	380	Translate and validate the EORTC IN-PATSAT32 for the Iranian general cancer population
Zhang et al., 2014 [201]	Chinese gastrointestinal cancer patients	106	Translate and validate the EORTC IN-PATSAT32 for the Chinese gastrointestinal cancer population
Zhang et al., 2015 [202]	Chinese cancer patients	302	Validate the EORTC IN-PATSAT32 for the Chinese general cancer population

Supplementary Table 7.12 Internal consistency (Cronbach Alpha's) of the IN-PATSAT32.

Reference	DrTech	DrInt	DrInfo	DrAva	NTech	NInt	NInfo
Arraras et al., 2009 [200]	0.91	0.94	0.95	0.89	0.97	0.93	0.98
Obtel et al., 2017 [203]	0.9	0.96	0.93	0.92	0.9	0.93	0.85
Pishkuhi et al., 2014 [199]	0.9	0.92	0.87	0.79	0.88	0.84	0.92
Zhang et al., 2014 [201]	0.91	0.91	0.93	0.93	0.87	0.88	0.87
Zhang et al., 2015 [202]	0.87	0.93	0.94	0.93	0.9	0.93	0.92

Reference	NAva	SInt	WT	HA	Rating	Quality
Arraras et al., 2009 [200]	0.77	0.79	0.87	0.36	?	Poor
Obtel et al., 2017 [203]	0.92	0.88	0.84	0.71	?	Poor
Pishkuhi et al., 2014 [199]	0.83	0.87	0.84	0.67	?	Fair
Zhang et al., 2014 [201]	0.81	0.84	0.81	0.74	?	Poor
Zhang et al., 2015 [Zhang2015]	0.85	0.88	0.86	0.86	?	Poor

IIEF: International Index of Erectile Function; EF: Erectile Function; OS: Overall Satisfaction; MCID: Minimal Clinically Important Difference * Due to an extremely small N

Supplementary Table 7.13. Test-retest reliability (correlation coefficients) of the IN-PATSAT32.

Reference	DrTech		DrInt	DrInfo		DrAva	NTech	NInt	NInfo
Obtel et al., 2017 [203]	0.88		0.91	0.91		0.64	0.89	0.73	0.86
Pishkuhi et al., 2014 [199]	0.88		0.87	0.92		0.87	0.91	0.94	0.96
Reference	**NAva**	**SInt**	**WT**	**HA**	**IE**	**HC**	**OA**	**Rating**	**Quality**
Obtel et al., 2017 [203]	0.87	0.82	0.7	0.75	0.84	0.73	0.67	?	Poor
Pishkuhi et al., 2014 [199]	0.91	0.93	0.9	0.86				?	Poor

FSFI: Female Sexual Function Index; DE: Desire; AR: Arousal; LU: Lubrication; OR: Orgasm; SA: Satisfaction; PA: Pain; BC: Breast Cancer

Appendices

Appendix A: Search string all PROMs

Search Terms

Embase.com

('Perceived Stress Scale'/de OR 'Insomnia Severity Index'/de OR 'International Index of Erectile Function'/de OR ((cancer NEAR/3 worr* NEAR/3 scale*) OR (patient NEAR/3 specifieke NEAR/3 klacht*) OR (insomni* NEAR/3 sever* NEAR/3 index*) OR (6-item NEAR/6 female NEAR/3 sexual* NEAR/3 function*) OR (5-item NEAR/6 erectile NEAR/3 function*) OR (sexual* NEAR/3 health NEAR/3 inventor* NEAR/3 men) OR (body NEAR/3 image NEAR/3 scal*) OR ((EORTC OR 'European Organization for Research and Treatment of Cancer') NEAR/6 (QLQ OR 'Quality of Life') NEAR/6 (PATSAT32 OR BR23 OR BR-23 OR CR-29 OR CR29 OR H&N25 OR HN25 OR HN-25)) OR (Caron NEAR/3 screening NEAR/3 questionnaire*) OR (Jong NEAR/3 Gierveld NEAR/3 loneliness) OR (7-item NEAR/3 dyadis NEAR/3 adjustment*) OR (vragenlijst NEAR/3 gezinskenmerken) OR (job NEAR/3 content* NEAR/3 questionnaire*) OR (vragenlijst NEAR/3 beleving NEAR/3 beoordeling NEAR/3 arbeid) OR (Alcohol NEAR/3 five-shot) OR (perceived NEAR/3 stress NEAR/3 scale*) OR (functional NEAR/3 assessment NEAR/3 cancer NEAR/3 therap* NEAR/3 endocrine) OR (breast NEAR/3 impact NEAR/3 treatment NEAR/3 scale*) OR (breast NEAR/3 reconstruction NEAR/3 satisfaction NEAR/3 questionnair*) OR (breast NEAR/3 cancer NEAR/3 patients NEAR/3 needs NEAR/3 questionnaire*) OR (stoma NEAR/3 quality NEAR/3 life NEAR/3 questionnaire*) OR (shoulder* NEAR/3 disabilit* NEAR/3 questionnaire*) OR (('CWS' OR 'SPK' OR 'FSFI-6' OR 'IIEF-5' OR 'CARON' OR 'JGLS' OR 'DAS-7' OR 'VGK-SF' OR 'JCQ' OR 'VBBA' OR 'A5S' OR 'FACT-ES' OR 'BITS' OR 'BRECON-31' OR 'BR-CNPQ' OR 'SDQ' OR 'stoma-QoL') NEAR/3 (assess* OR score* OR scale* OR questionnaire* OR inventor* OR measure*))):ab,ti) AND (neoplasm/exp OR (neoplas* OR cancer* OR oncolog* OR tumor* OR tumour OR carcino*):ab,ti) AND ('validation study'/de OR 'reproducibility'/de OR 'psychometry'/de OR 'observer variation'/de OR 'discriminant analysis'/de OR 'correlation coefficient'/de OR reliability/de OR 'sensitivity and specificity'/de OR validity/exp OR 'sensitivity analysis'/de OR 'internal consistency'/de OR 'confidence interval'/de OR (psychometr* OR reproducib* OR clinimetr* OR clinometr* OR observer-varia* OR reliab* OR valid* OR coefficient OR interna*-consisten* OR (cronbach* NEAR/3 (alpha OR alphas)) OR (item* NEXT/1 (correlation* OR selection* OR reduction*)) OR agreement OR precision OR imprecision OR precise-value* OR test*-retest* OR (test NEAR/3 retest) OR (reliab* NEAR/3 (test OR retest)) OR stability OR interrater OR inter-rater OR intrarater OR intra-rater OR intertester OR inter-tester OR intratester OR intra-tester OR interobserver OR inter-

observer OR intraobserver OR intra-observer OR intertechnician OR inter-technician OR intratechnician OR intra-technician OR interexaminer OR inter-examiner OR intraexaminer OR intra-examiner OR interassay OR inter-assay OR intraassay OR intra-assay OR interindividual OR inter-individual OR intraindividual OR intra-individual OR interparticipant OR inter-participant OR intraparticipant OR intra-participant OR kappa OR kappa-s OR kappas OR (coefficient* NEAR/3 variation*) OR repeatab* OR ((replicab* OR repeat*) NEAR/3 (measure OR measures OR findings OR result OR results OR test OR tests)) OR generaliza* OR generalisa* OR concordance OR (intraclass NEAR/3 correlation*) OR discriminative OR 'known group' OR (factor* NEAR/3 (analys* OR structure*)) OR dimensionality OR subscale* OR (multitrait NEAR/3 scaling) OR item-discriminant* OR (interscale NEAR/3 correlat*) OR ((error OR errors) NEAR/3 (measure* OR correlat* OR evaluat* OR accuracy OR accurate OR precision OR mean)) OR ((individual OR interval OR rate OR analy*) NEAR/3 variabilit*) OR (uncertaint* NEAR/3 (measure*)) OR (error NEAR/3 measure*) OR sensitiv* OR responsive* OR (limit NEAR/3 detection) OR (minimal* NEAR/3 detectab*) OR interpretab* OR (small* NEAR/3 (real OR detectable) NEAR/3 (change OR difference)) OR (meaningful* NEAR/3 change*) OR (minimal* NEAR/3 (important OR detectab* OR real) NEAR/3 (change* OR difference)) OR ((ceiling OR floor) NEXT/1 effect*) OR 'Item response model' OR IRT OR Rasch OR 'Differential item functioning' OR DIF OR 'computer adaptive testing' OR 'item bank' OR 'cross-cultural equivalence' OR (confidence* NEAR/3 interval*)):ab,ti)

Medline Ovid

(((cancer ADJ3 worr* ADJ3 scale*) OR (patient ADJ3 specifieke ADJ3 klacht*) OR (insomni* ADJ3 sever* ADJ3 index*) OR (6-item ADJ6 female ADJ3 sexual* ADJ3 function*) OR (5-item ADJ6 erectile ADJ3 function*) OR (sexual* ADJ3 health ADJ3 inventor* ADJ3 men) OR (body ADJ3 image ADJ3 scal*) OR ((EORTC OR "European Organization for Research and Treatment of Cancer") ADJ6 (QLQ OR "Quality of Life") ADJ6 (PATSAT32 OR BR23 OR BR-23 OR CR-29 OR CR29 OR H&N25 OR HN25 OR HN-25)) OR (Caron ADJ3 screening ADJ3 questionnaire*) OR (Jong ADJ3 Gierveld ADJ3 loneliness) OR (7-item ADJ3 dyadis ADJ3 adjustment*) OR (vragenlijst ADJ3 gezinskenmerken) OR (job ADJ3 content* ADJ3 questionnaire*) OR (vragenlijst ADJ3 beleving ADJ3 beoordeling ADJ3 arbeid) OR (Alcohol ADJ3 five-shot) OR (perceived ADJ3 stress ADJ3 scale*) OR (functional ADJ3 assessment ADJ3 cancer ADJ3 therap* ADJ3 endocrine) OR (breast ADJ3 impact ADJ3 treatment ADJ3 scale*) OR (breast ADJ3 reconstruction ADJ3 satisfaction ADJ3 questionnair*) OR (breast ADJ3 cancer ADJ3 patients ADJ3 needs ADJ3 questionnaire*) OR (stoma ADJ3 quality ADJ3 life ADJ3 questionnaire*) OR (shoulder* ADJ3 disabilit* ADJ3 questionnaire*) OR (("CWS" OR "SPK" OR "FSFI-6" OR "IIEF-5" OR "CARON" OR "JGLS" OR "DAS-7" OR "VGK-SF" OR "JCQ" OR "VBBA" OR "A5S" OR

"FACT-ES" OR "BITS" OR "BRECON-31" OR "BR-CNPQ" OR "SDQ" OR "stoma-QoL") ADJ3 (assess* OR score* OR scale* OR questionnaire* OR inventor* OR measure*))).ab,ti.) AND (neoplasm/ OR (neoplas* OR cancer* OR oncolog* OR tumor* OR tumour OR carcino*).ab,ti.) AND (exp "Validation Studies"/ OR exp "reproducibility of results"/ OR exp "psychometrics"/ OR exp "observer variation"/ OR exp "discriminant analysis"/ OR exp "Sensitivity and Specificity"/ OR "Confidence Intervals"/ OR (psychometr* OR reproducib* OR clinimetr* OR clinometr* OR observer-varia* OR reliab* OR valid* OR coefficient OR interna*-consisten* OR (cronbach* ADJ3 (alpha OR alphas)) OR (item* ADJ (correlation* OR selection* OR reduction*)) OR agreement OR precision OR imprecision OR precise-value* OR test*-retest* OR (test ADJ3 retest) OR (reliab* ADJ3 (test OR retest)) OR stability OR interrater OR inter-rater OR intrarater OR intra-rater OR intertester OR inter-tester OR intratester OR intra-tester OR interobserver OR inter-observer OR intraobserver OR intra-observer OR intertechnician OR inter-technician OR intratechnician OR intra-technician OR interexaminer OR inter-examiner OR intraexaminer OR intra-examiner OR interassay OR inter-assay OR intraassay OR intra-assay OR interindividual OR inter-individual OR intraindividual OR intra-individual OR interparticipant OR inter-participant OR intraparticipant OR intra-participant OR kappa OR kappa-s OR kappas OR (coefficient* ADJ3 variation*) OR repeatab* OR ((replicab* OR repeat*) ADJ3 (measure OR measures OR findings OR result OR results OR test OR tests)) OR generaliza* OR generalisa* OR concordance OR (intraclass ADJ3 correlation*) OR discriminative OR "known group" OR (factor* ADJ3 (analys* OR structure*)) OR dimensionality OR subscale* OR (multitrait ADJ3 scaling) OR item-discriminant* OR (interscale ADJ3 correlat*) OR ((error OR errors) ADJ3 (measure* OR correlat* OR evaluat* OR accuracy OR accurate OR precision OR mean)) OR ((individual OR interval OR rate OR analy*) ADJ3 variabilit*) OR (uncertaint* ADJ3 (measure*)) OR (error ADJ3 measure*) OR sensitiv* OR responsive* OR (limit ADJ3 detection) OR (minimal* ADJ3 detectab*) OR interpretab* OR (small* ADJ3 (real OR detectable) ADJ3 (change OR difference)) OR (meaningful* ADJ3 change*) OR (minimal* ADJ3 (important OR detectab* OR real) ADJ3 (change* OR difference)) OR ((ceiling OR floor) ADJ effect*) OR "Item response model" OR IRT OR Rasch OR "Differential item functioning" OR DIF OR "computer adaptive testing" OR "item bank" OR "cross-cultural equivalence" OR (confidence* ADJ3 interval*)).ab,ti.)

PsycINFO Ovid

(((cancer ADJ3 worr* ADJ3 scale*) OR (patient ADJ3 specifieke ADJ3 klacht*) OR (insomni* ADJ3 sever* ADJ3 index*) OR (6-item ADJ6 female ADJ3 sexual* ADJ3 function*) OR (5-item ADJ6 erectile ADJ3 function*) OR (sexual* ADJ3 health ADJ3 inventor* ADJ3 men) OR (body ADJ3 image ADJ3 scal*) OR ((EORTC OR "European Organization for Research and Treatment of Cancer") ADJ6 (QLQ OR "Quality of

Life") ADJ6 (PATSAT32 OR BR23 OR BR-23 OR CR-29 OR CR29 OR H&N25 OR HN25 OR HN-25)) OR (Caron ADJ3 screening ADJ3 questionnaire*) OR (Jong ADJ3 Gierveld ADJ3 loneliness) OR (7-item ADJ3 dyadis ADJ3 adjustment*) OR (vragenlijst ADJ3 gezinskenmerken) OR (job ADJ3 content* ADJ3 questionnaire*) OR (vragenlijst ADJ3 beleving ADJ3 beoordeling ADJ3 arbeid) OR (Alcohol ADJ3 five-shot) OR (perceived ADJ3 stress ADJ3 scale*) OR (functional ADJ3 assessment ADJ3 cancer ADJ3 therap* ADJ3 endocrine) OR (breast ADJ3 impact ADJ3 treatment ADJ3 scale*) OR (breast ADJ3 reconstruction ADJ3 satisfaction ADJ3 questionnair*) OR (breast ADJ3 cancer ADJ3 patients ADJ3 needs ADJ3 questionnaire*) OR (stoma ADJ3 quality ADJ3 life ADJ3 questionnaire*) OR (shoulder* ADJ3 disabilit* ADJ3 questionnaire*) OR (("CWS" OR "SPK" OR "FSFI-6" OR "IIEF-5" OR "CARON" OR "JGLS" OR "DAS-7" OR "VGK-SF" OR "JCQ" OR "VBBA" OR "A5S" OR "FACT-ES" OR "BITS" OR "BRECON-31" OR "BR-CNPQ" OR "SDQ" OR "stoma-QoL") ADJ3 (assess* OR score* OR scale* OR questionnaire* OR inventor* OR measure*))).ab,ti.) AND (neoplasm/ OR (neoplas* OR cancer* OR oncolog* OR tumor* OR tumour OR carcino*).ab,ti.) AND (exp "Test Validity"/ OR exp "Test Reliability"/ OR exp "psychometrics"/ OR exp "Interrater Reliability"/ OR exp OR (psychometr* OR reproducib* OR clinimetr* OR clinometr* OR observer-varia* OR reliab* OR valid* OR coefficient OR interna*-consisten* OR (cronbach* ADJ3 (alpha OR alphas)) OR (item* ADJ (correlation* OR selection* OR reduction*)) OR agreement OR precision OR imprecision OR precise-value* OR test*-retest* OR (test ADJ3 retest) OR (reliab* ADJ3 (test OR retest)) OR stability OR interrater OR inter-rater OR intrarater OR intra-rater OR intertester OR inter-tester OR intratester OR intra-tester OR interobserver OR inter-observer OR intraobserver OR intra-observer OR intertechnician OR inter-technician OR intratechnician OR intra-technician OR interexaminer OR inter-examiner OR intraexaminer OR intra-examiner OR interassay OR inter-assay OR intraassay OR intra-assay OR interindividual OR inter-individual OR intraindividual OR intra-individual OR interparticipant OR inter-participant OR intraparticipant OR intra-participant OR kappa OR kappa-s OR kappas OR (coefficient* ADJ3 variation*) OR repeatab* OR ((replicab* OR repeat*) ADJ3 (measure OR measures OR findings OR result OR results OR test OR tests)) OR generaliza* OR generalisa* OR concordance OR (intraclass ADJ3 correlation*) OR discriminative OR "known group" OR (factor* ADJ3 (analys* OR structure*)) OR dimensionality OR subscale* OR (multitrait ADJ3 scaling) OR item-discriminant* OR (interscale ADJ3 correlat*) OR ((error OR errors) ADJ3 (measure* OR correlat* OR evaluat* OR accuracy OR accurate OR precision OR mean)) OR ((individual OR interval OR rate OR analy*) ADJ3 variabilit*) OR (uncertaint* ADJ3 (measure*)) OR (error ADJ3 measure*) OR sensitiv* OR responsive* OR (limit ADJ3 detection) OR (minimal* ADJ3 detectab*) OR interpretab* OR (small* ADJ3 (real OR detectable) ADJ3 (change OR difference)) OR (meaningful* ADJ3 change*) OR (minimal* ADJ3

(important OR detectab* OR real) ADJ3 (change* OR difference)) OR ((ceiling OR floor) ADJ effect*) OR "Item response model" OR IRT OR Rasch OR "Differential item functioning" OR DIF OR "computer adaptive testing" OR "item bank" OR "cross-cultural equivalence" OR (confidence* ADJ3 interval*)).ab,ti.)

Web of science

TS=((((cancer NEAR/3 worr* NEAR/3 scale*) OR (patient NEAR/3 specifieke NEAR/3 klacht*) OR (insomni* NEAR/3 sever* NEAR/3 index*) OR (6-item NEAR/6 female NEAR/3 sexual* NEAR/3 function*) OR (5-item NEAR/6 erectile NEAR/3 function*) OR (sexual* NEAR/3 health NEAR/3 inventor* NEAR/3 men) OR (body NEAR/3 image NEAR/3 scal*) OR ((EORTC OR "European Organization for Research and Treatment of Cancer") NEAR/6 (QLQ OR "Quality of Life") NEAR/6 (PATSAT32 OR BR23 OR BR-23 OR CR-29 OR CR29 OR H&N25 OR HN25 OR HN-25)) OR (Caron NEAR/3 screening NEAR/3 questionnaire*) OR (Jong NEAR/3 Gierveld NEAR/3 loneliness) OR (7-item NEAR/3 dyadis NEAR/3 adjustment*) OR (vragenlijst NEAR/3 gezinskenmerken) OR (job NEAR/3 content* NEAR/3 questionnaire*) OR (vragenlijst NEAR/3 beleving NEAR/3 beoordeling NEAR/3 arbeid) OR (Alcohol NEAR/3 five-shot) OR (perceived NEAR/3 stress NEAR/3 scale*) OR (functional NEAR/3 assessment NEAR/3 cancer NEAR/3 therap* NEAR/3 endocrine) OR (breast NEAR/3 impact NEAR/3 treatment NEAR/3 scale*) OR (breast NEAR/3 reconstruction NEAR/3 satisfaction NEAR/3 questionnair*) OR (breast NEAR/3 cancer NEAR/3 patients NEAR/3 needs NEAR/3 questionnaire*) OR (stoma NEAR/3 quality NEAR/3 life NEAR/3 questionnaire*) OR (shoulder* NEAR/3 disabilit* NEAR/3 questionnaire*) OR (("CWS" OR "SPK" OR "FSFI-6" OR "IIEF-5" OR "CARON" OR "JGLS" OR "DAS-7" OR "VGK-SF" OR "JCQ" OR "VBBA" OR "A5S" OR "FACT-ES" OR "BITS" OR "BRECON-31" OR "BR-CNPQ" OR "SDQ" OR "stoma-QoL") NEAR/3 (assess* OR score* OR scale* OR questionnaire* OR inventor* OR measure*)))) AND (neoplasm/exp OR (neoplas* OR cancer* OR oncolog* OR tumor* OR tumour OR carcino*)) AND ((psychometr* OR reproducib* OR clinimetr* OR clinometr* OR observer-varia* OR reliab* OR valid* OR coefficient OR interna*-consisten* OR (cronbach* NEAR/3 (alpha OR alphas)) OR (item* NEAR/1 (correlation* OR selection* OR reduction*)) OR agreement OR precision OR imprecision OR precise-value* OR test*-retest* OR (test NEAR/3 retest) OR (reliab* NEAR/3 (test OR retest)) OR stability OR interrater OR inter-rater OR intrarater OR intra-rater OR intertester OR inter-tester OR intratester OR intra-tester OR interobserver OR inter-observer OR intraobserver OR intra-observer OR intertechnician OR inter-technician OR intratechnician OR intra-technician OR interexaminer OR inter-examiner OR intraexaminer OR intra-examiner OR interassay OR inter-assay OR intraassay OR intra-assay OR interindividual OR inter-individual OR intraindividual OR intra-individual OR interparticipant OR inter-participant OR

intraparticipant OR intra-participant OR kappa OR kappa-s OR kappas OR (coefficient* NEAR/3 variation*) OR repeatab* OR ((replicab* OR repeat*) NEAR/3 (measure OR measures OR findings OR result OR results OR test OR tests)) OR generaliza* OR generalisa* OR concordance OR (intraclass NEAR/3 correlation*) OR discriminative OR "known group" OR (factor* NEAR/3 (analys* OR structure*)) OR dimensionality OR subscale* OR (multitrait NEAR/3 scaling) OR item-discriminant* OR (interscale NEAR/3 correlat*) OR ((error OR errors) NEAR/3 (measure* OR correlat* OR evaluat* OR accuracy OR accurate OR precision OR mean)) OR ((individual OR interval OR rate OR analy*) NEAR/3 variabilit*) OR (uncertaint* NEAR/3 (measure*)) OR (error NEAR/3 measure*) OR sensitiv* OR responsive* OR (limit NEAR/3 detection) OR (minimal* NEAR/3 detectab*) OR interpretab* OR (small* NEAR/3 (real OR detectable) NEAR/3 (change OR difference)) OR (meaningful* NEAR/3 change*) OR (minimal* NEAR/3 (important OR detectab* OR real) NEAR/3 (change* OR difference)) OR ((ceiling OR floor) NEAR/1 effect*) OR "Item response model" OR IRT OR Rasch OR "Differential item functioning" OR DIF OR "computer adaptive testing" OR "item bank" OR "cross-cultural equivalence" OR (confidence* NEAR/3 interval*))))

Appendix B: Methodological Quality Assessments EORTC IN-PAT-SAT32

Table B.1. Quality Assessment Structural Validity.

Reference	Overall score (lowest grade)	Is the scale based on a reflective model?	Was the percentage of missing items given?	Was there a description of how missing items were handled?	Was the sample size included in the analysis adequate?	Were there any important flaws in the design or methods of the study?	for CTT: Was exploratory or confirmatory factor analysis performed?
Arraras (2009) [200]	Poor	Yes	Good	Excellent	Poor	Fair	Poor
Hjörleifsdóttir (2010) [198]	Good	Yes	Excellent	Excellent	Good	Excellent	Good
Obtel (2017) [203]	Poor	Yes	Good	Fair	Poor	Fair	Poor
Pishkuhi (2014) [199]	Fair	Yes	Good	Fair	Excellent	Excellent	Good
Zhang (2014) [201]	Poor	Yes	Good	Fair	Poor	Fair	Poor
Zhang (2015) [202]	Poor	Yes	Good	Fair	Excellent	Fair	Poor

Table B.2. Quality Assessment Internal Consistency.

Reference	Overall score (lowest grade)	Is the scale based on a reflective model?	Was the percentage of missing items given?	Was there a description of how missing items were handled?	Was the sample size included in the internal consistency analysis adequate?
Arraras (2009) [200]	Poor	Yes	Good	Excellent	Good
Hjörleifsdóttir (2010) [198]	Good	Yes	Excellent	Excellent	Excellent
Obtel (2017) [203]	Poor	Yes	Good	Fair	Excellent
Pishkuhi (2014) [199]	Fair	Yes	Good	Fair	Excellent
Zhang (2014) [201]	Poor	Yes	Good	Fair	Excellent
Zhang (2015) [202]	Poor	Yes	Good	Fair	Excellent

Reference	Was the unidimensionality of the scale checked?	Was the sample size included in the unidimensionality analysis adequate?	Was an internal consistency statistic calculated for each (unidimensional) (sub)scale separately?	Were there any important flaws in the design or methods of the study?	For Classical Test Theory (CTT), continuous scores: Was Cronbach's alpha calculated?
Arraras (2009) [200]	Poor	Excellent	Excellent	Excellent	Excellent
Hjörleifsdóttir (2010) [198]	Excellent	Good	Excellent	Excellent	Excellent
Obtel (2017) [203]	Poor	Poor	Excellent	Excellent	Excellent
Pishkuhi (2014) [199]	Excellent	Excellent	Excellent	Excellent	Excellent
Zhang (2014) [201]	Poor	Poor	Excellent	Excellent	Excellent
Zhang (2015) [202]	Poor	Excellent	Excellent	Excellent	Excellent

Table B.3. Quality Assessment Reliability.

Reference	Overall score (lowest grade)	Was the percentage of missing items given?	Was there a description of how missing items were handled?	Was the sample size included in the analysis adequate?	Were at least two measurements available?	Were the administrations independent?	Was the time interval stated?
Obtel (2017) [203]	Fair	Good	Fair	Excellent	Excellent	Excellent	Excellent
Pishkuhi (2014) [199]	Fair	Good	Fair	Good	Excellent	Excellent	Excellent

Reference	Were patients stable in the interim period on the construct to be measured?	Was the time interval appropriate?	Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Were there any important flaws in the design or methods of the study?	for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?
Obtel (2017) [203]	Good	Fair	Good	Excellent	Good
Pishkuhi (2014) [199]	Good	Excellent	Good	Excellent	Fair

Table B.4. Quality Assessment Hypothesis Testing.

Reference	Overall score (lowest grade)	Was the percentage of missing items given?	Was there a description of how missing items were handled?	Was the sample size included in the analysis adequate?	Were hypotheses regarding correlations or mean differences formulated a priori (i.e. before data collection)?	Was the expected direction of correlations or mean differences included in the hypotheses?
Aboshaiqah (2016) [208]	Poor	Good	Fair	Excellent	Poor	Good
Arraras (2009) [200]	Fair	Good	Excellent	Good	Fair	Good
Arraras (2010) [206]	Fair	Good	Fair	Excellent	Excellent	Excellent
Asadi-lari (2015) [205]	Good	Excellent	Excellent	Excellent	Excellent	Excellent
Pishkuhi (2014) [199]	Fair	Good	Fair	Excellent	Good	Good
Zhang (2014) [201]	Poor	Good	Good	Excellent	Poor	Good
Zhang (2015) [202]	Poor	Good	Good	Excellent	Poor	Good

Reference	Was the expected absolute or relative magnitude of correlations or mean differences included in the hypotheses?	for convergent validity: Was an adequate description provided of the comparator instrument(s)?	for convergent validity: Were the measurement properties of the comparator instrument(s) adequately described?	Were there any important flaws in the design or methods of the study?	Were design and statistical methods adequate for the hypotheses to be tested?
Aboshaiqah (2016) [208]	Good	Good	Poor	Excellent	Good
Arraras (2009) [200]	Good	Excellent	Fair	Excellent	Good
Arraras (2010) [206]	Excellent	Excellent	Excellent	Excellent	Good
Asadi-lari (2015) [205]	Excellent	Excellent	Excellent	Excellent	Good
Pishkuhi (2014) [199]	Excellent	n/a	n/a	Excellent	Good
Zhang (2014) [201]	Good	n/a	n/a	Excellent	Good
Zhang (2015) [202]	Good	n/a	n/a	Excellent	Good

Table B.5. Quality Assessment Measurement Error.

Reference	Overall score (lowest grade)	Was the percentage of missing items given?	Was there a description of how missing items were handled?	Was the sample size included in the analysis adequate?	Were at least two measurements available?	Were the administrations independent?	Was the time interval stated?
Arraras (2009) [200]	Poor	Good	Excellent	Good	Poor	n/a	n/a
Obtel (2017) [203]	Fair	Good	Fair	Excellent	Excellent	Excellent	Excellent
Pishkuhi (2014) [199]	Fair	Good	Fair	Excellent	Excellent	Excellent	Excellent
Zhang (2014) [201]	Poor	Good	Fair	Excellent	Poor	n/a	n/a
Zhang (2015) [202]	Poor	Good	Fair	Excellent	Poor	n/a	n/a

Reference	Were patients stable in the interim period on the construct to be measured?	Was the time interval appropriate?	Were the test conditions similar for both measurements? e.g. type of administration, environment, instructions	Were there any important flaws in the design or methods of the study?	For CTT: Was the Standard Error of Measurement (SEM), Smallest Detectable Change (SDC) or Limits of Agreement (LoA) calculated?
Arraras (2009) [200]	n/a	n/a	n/a	Excellent	Poor
Obtel (2017) [203]	Good	Fair	Good	Excellent	Good
Pishkuhi (2014) [199]	Good	Excellent	Good	Excellent	Good
Zhang (2014) [201]	n/a	n/a	n/a	Excellent	Poor
Zhang (2015) [202]	n/a	n/a	n/a	Excellent	Poor

Appendix C: Search Terms IIEF-specific

Search Terms

Embase.com

((5-item NEAR/6 erectile NEAR/3 function*) OR (('IIEF') NEAR/3 (assess* OR score* OR scale* OR questionnaire* OR inventor* OR measure*)):ab,ti) AND ('validation study'/de OR 'reproducibility'/de OR 'psychometry'/de OR 'observer variation'/de OR 'discriminant analysis'/de OR 'correlation coefficient'/de OR reliability/de OR 'sensitivity and specificity'/de OR validity/exp OR 'sensitivity analysis'/de OR 'internal consistency'/de OR 'confidence interval'/de OR (psychometr* OR reproducib* OR clinimetr* OR clinometr* OR observer-varia* OR reliab* OR valid* OR coefficient OR interna*-consisten* OR (cronbach* NEAR/3 (alpha OR alphas)) OR (item* NEXT/1 (correlation* OR selection* OR reduction*)) OR agreement OR precision OR imprecision OR precise-value* OR test*-retest* OR (test NEAR/3 retest) OR (reliab* NEAR/3 (test OR retest)) OR stability OR interrater OR inter-rater OR intrarater OR intra-rater OR intertester OR inter-tester OR intratester OR intra-tester OR interobserver OR inter-observer OR intraobserver OR intra-observer OR intertechnician OR inter-technician OR intratechnician OR intra-technician OR interexaminer OR inter-examiner OR intraexaminer OR intra-examiner OR interassay OR inter-assay OR intraassay OR intra-assay OR interindividual OR inter-individual OR intraindividual OR intra-individual OR interparticipant OR inter-participant OR intraparticipant OR intra-participant OR kappa OR kappa-s OR kappas OR (coefficient* NEAR/3 variation*) OR repeatab* OR ((replicab* OR repeat*) NEAR/3 (measure OR measures OR findings OR result OR results OR test OR tests)) OR generaliza* OR generalisa* OR concordance OR (intraclass NEAR/3 correlation*) OR discriminative OR 'known group' OR (factor* NEAR/3 (analys* OR structure*)) OR dimensionality OR subscale* OR (multitrait NEAR/3 scaling) OR item-discriminant* OR (interscale NEAR/3 correlat*) OR ((error OR errors) NEAR/3 (measure* OR correlat* OR evaluat* OR accuracy OR accurate OR precision OR mean)) OR ((individual OR interval OR rate OR analy*) NEAR/3 variabilit*) OR (uncertaint* NEAR/3 (measure*)) OR (error NEAR/3 measure*) OR sensitiv* OR responsive* OR (limit NEAR/3 detection) OR (minimal* NEAR/3 detectab*) OR interpretab* OR (small* NEAR/3 (real OR detectable) NEAR/3 (change OR difference)) OR (meaningful* NEAR/3 change*) OR (minimal* NEAR/3 (important OR detectab* OR real) NEAR/3 (change* OR difference)) OR ((ceiling OR floor) NEXT/1 effect*) OR 'Item response model' OR IRT OR Rasch OR 'Differential item functioning' OR DIF OR 'computer adaptive testing' OR 'item bank' OR 'cross-cultural equivalence' OR (confidence* NEAR/3 interval*)):ab,ti)

Medline Ovid

(((5-item ADJ6 erectile ADJ3 function*) OR (("IIEF") ADJ3 (assess* OR score* OR scale* OR questionnaire* OR inventor* OR measure*))).ab,ti.) AND (exp "Validation Studies"/ OR exp "reproducibility of results"/ OR exp "psychometrics"/ OR exp "observer variation"/ OR exp "discriminant analysis"/ OR exp "Sensitivity and Specificity"/ OR "Confidence Intervals"/ OR (psychometr* OR reproducib* OR clinimetr* OR clinometr* OR observer-varia* OR reliab* OR valid* OR coefficient OR interna*-consisten* OR (cronbach* ADJ3 (alpha OR alphas)) OR (item* ADJ (correlation* OR selection* OR reduction*)) OR agreement OR precision OR imprecision OR precise-value* OR test*-retest* OR (test ADJ3 retest) OR (reliab* ADJ3 (test OR retest)) OR stability OR interrater OR inter-rater OR intrarater OR intra-rater OR intertester OR inter-tester OR intratester OR intra-tester OR interobserver OR inter-observer OR intraobserver OR intra-observer OR intertechnician OR inter-technician OR intratechnician OR intra-technician OR interexaminer OR inter-examiner OR intraexaminer OR intra-examiner OR interassay OR inter-assay OR intraassay OR intra-assay OR interindividual OR inter-individual OR intraindividual OR intra-individual OR interparticipant OR inter-participant OR intraparticipant OR intra-participant OR kappa OR kappa-s OR kappas OR (coefficient* ADJ3 variation*) OR repeatab* OR ((replicab* OR repeat*) ADJ3 (measure OR measures OR findings OR result OR results OR test OR tests)) OR generaliza* OR generalisa* OR concordance OR (intraclass ADJ3 correlation*) OR discriminative OR "known group" OR (factor* ADJ3 (analys* OR structure*)) OR dimensionality OR subscale* OR (multitrait ADJ3 scaling) OR item-discriminant* OR (interscale ADJ3 correlat*) OR ((error OR errors) ADJ3 (measure* OR correlat* OR evaluat* OR accuracy OR accurate OR precision OR mean)) OR ((individual OR interval OR rate OR analy*) ADJ3 variabilit*) OR (uncertaint* ADJ3 (measure*)) OR (error ADJ3 measure*) OR sensitiv* OR responsive* OR (limit ADJ3 detection) OR (minimal* ADJ3 detectab*) OR interpretab* OR (small* ADJ3 (real OR detectable) ADJ3 (change OR difference)) OR (meaningful* ADJ3 change*) OR (minimal* ADJ3 (important OR detectab* OR real) ADJ3 (change* OR difference)) OR ((ceiling OR floor) ADJ effect*) OR "Item response model" OR IRT OR Rasch OR "Differential item functioning" OR DIF OR "computer adaptive testing" OR "item bank" OR "cross-cultural equivalence" OR (confidence* ADJ3 interval*)).ab,ti.)

Web of science

TS=((((5-item NEAR/6 erectile NEAR/3 function*) OR (("IIEF") NEAR/3 (assess* OR score* OR scale* OR questionnaire* OR inventor* OR measure*)))) AND ((psychometr* OR reproducib* OR clinimetr* OR clinometr* OR observer-varia* OR reliab* OR valid* OR coefficient OR interna*-consisten* OR (cronbach* NEAR/3 (alpha OR alphas)) OR (item* NEAR/1 (correlation* OR selection* OR reduction*))

OR agreement OR precision OR imprecision OR precise-value* OR test*-retest* OR (test NEAR/3 retest) OR (reliab* NEAR/3 (test OR retest)) OR stability OR interrater OR inter-rater OR intrarater OR intra-rater OR intertester OR inter-tester OR intratester OR intra-tester OR interobserver OR inter-observer OR intraobserver OR intra-observer OR intertechnician OR inter-technician OR intratechnician OR intra-technician OR interexaminer OR inter-examiner OR intraexaminer OR intra-examiner OR interassay OR inter-assay OR intraassay OR intra-assay OR interindividual OR inter-individual OR intraindividual OR intra-individual OR interparticipant OR inter-participant OR intraparticipant OR intra-participant OR kappa OR kappa-s OR kappas OR (coefficient* NEAR/3 variation*) OR repeatab* OR ((replicab* OR repeat*) NEAR/3 (measure OR measures OR findings OR result OR results OR test OR tests)) OR generaliza* OR generalisa* OR concordance OR (intraclass NEAR/3 correlation*) OR discriminative OR "known group" OR (factor* NEAR/3 (analys* OR structure*)) OR dimensionality OR subscale* OR (multitrait NEAR/3 scaling) OR item-discriminant* OR (interscale NEAR/3 correlat*) OR ((error OR errors) NEAR/3 (measure* OR correlat* OR evaluat* OR accuracy OR accurate OR precision OR mean)) OR ((individual OR interval OR rate OR analy*) NEAR/3 variabilit*) OR (uncertaint* NEAR/3 (measure*)) OR (error NEAR/3 measure*) OR sensitiv* OR responsive* OR (limit NEAR/3 detection) OR (minimal* NEAR/3 detectab*) OR interpretab* OR (small* NEAR/3 (real OR detectable) NEAR/3 (change OR difference)) OR (meaningful* NEAR/3 change*) OR (minimal* NEAR/3 (important OR detectab* OR real) NEAR/3 (change* OR difference)) OR ((ceiling OR floor) NEAR/1 effect*) OR "Item response model" OR IRT OR Rasch OR "Differential item functioning" OR DIF OR "computer adaptive testing" OR "item bank" OR "cross-cultural equivalence" OR (confidence* NEAR/3 interval*))))

Appendix D: Methodological Quality Assessments IIEF

Table D.1. Quality Assessment Structural Validity.

Reference	Overall score (lowest grade)	Does the scale consist of effect indicators, i.e. is it based on a reflective model?	for CTT: Was exploratory or confirmatory factor analysis performed?	for IRT: does the chosen model fit to the research question?	Was the sample size included in the analysis adequate?	Were there any important flaws in the design or methods of the study?
Bushmakin (2014) [68]	Very good	Yes	Very good	n/a	Very good	Very good
Coyne (2010) [72]	Adequate	Yes	Adequate	n/a	Very good	Very good
Gonzáles (2013) [76]	Doubtful	Yes	Adequate	n/a	Doubtful	Very good
Kriston (2008) [78]	Very good	Yes	Very good	n/a	Very good	Very good
Lim (2003) [102]	Adequate	Yes	Adequate	n/a	Very good	Very good
Quinta Gomes (2012) [87]	Doubtful	Yes	Adequate	n/a	Very good	Doubtful
Rosen (1997) [52]	Adequate	Yes	Adequate	n/a	Very good	Very good
Wiltink (2003) [94]	Adequate	Yes	Adequate	n/a	Very good	Very good
Lin (2016) [96]	Very good	Yes	n/a	Very good	Very good	Very good

Table D.2. Quality Assessment Internal Consistency.

Reference	Overall score (lowest grade)	Does the scale consist of effect indicators, i.e. is it based on a reflective model?	Was an internal consistency statistic calculated for each unidimensional (sub) scale separately?	For continuous scores: Was Cronbach's alpha or omega calculated?	Were there any important flaws in the design or methods of the study?
Bayraktar (2012) [66]	Very good	Yes	Very good	Very good	Very good
Bayraktar (2013) [67]	Inadequate	Yes	Inadequate	Very good	Very good
Coyne (2010) [72]	Very good	Yes	Very good	Very good	Very good
Dargis (2013) [101]	Very good	Yes	Very good	Very good	Very good
Gonzáles (2013) [76]	Adequate	Yes	Very good	Very good	Adequate
Kriston (2008) [78]	Very good	Yes	Very good	Very good	Very good
Lim (2003) [102]	Very good	Yes	Very good	Very good	Very good
Mahmood (2012) [97]	Inadequate	Yes	Very good	Very good	Inadequate

Reference	Overall score (lowest grade)	Does the scale consist of effect indicators, i.e. is it based on a reflective model?	Was an internal consistency statistic calculated for each unidimensional (sub) scale separately?	For continuous scores: Was Cronbach's alpha or omega calculated?	Were there any important flaws in the design or methods of the study?
Quek (2002) [86]	Inadequate	Yes	Very good	Very good	Inadequate
Quinta Gomes (2012) [87]	Very good	Yes	Very good	Very good	Very good
Rosen (1997) [52]	Very good	Yes	Very good	Very good	Very good
Rubio-Aurioles (2009) [89]	Very good	Yes	Very good	Very good	Very good
Tang (2015) [98]	Very good	Yes	Very good	Very good	Very good
Utomo (2015) [99]	Very good	Yes	Very good	Very good	Very good
Wiltink (2003) [94]	Adequate	Yes	Very good	Very good	Adequate
Nimbi (2018) [81]	Adequate	Yes	Very good	Very good	Adequate
Pascoal (2017) [85]	Inadequate	Yes	Inadequate	Very good	Very good
Tang (2018) [92]	Inadequate	Yes	Inadequate	Very good	Very good

Table D.3. Quality Assessment Reliability.

Reference	Overall score (lowest grade)	Were patients stable in the interim period on the construct to be measured?	Was the time interval appropriate?	Were the test conditions similar for both measurements?	for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	Were there any important flaws in the design or methods of the study?
Bayraktar (2012) [66]	Inadequate	Very good	Very good	Inadequate	Doubtful	Very good
Bayraktar (2013) [67]	Doubtful	Very good	Very good	Very good	Doubtful	Very good
Gonzáles (2013) [76]	Doubtful	Adequate	Very good	Adequate	Adequate	Doubtful
Lim (2003) [102]	Doubtful	Adequate	Doubtful	Adequate	Adequate	Very good
Mahmood (2012) [97]	Inadequate	Adequate	Doubtful	Inadequate	Adequate	Very good
Quek (2002) [86]	Inadequate	Very good	Doubtful	Adequate	Adequate	Inadequate
Quinta Gomes (2012) [87]	Doubtful	Adequate	Very good	Adequate	Doubtful	Very good
Rosen (1997) [52]	Doubtful	Adequate	Very good	Adequate	Doubtful	Very good
Serefoglu (2008) [91]	Doubtful	Adequate	Doubtful	Adequate	Doubtful	Very good
Utomo (2015) [99]	Adequate	Very good	Very good	Adequate	Adequate	Very good

Table D.4. Quality Assessment Measurement Error.

Reference	Overall score (lowest grade)	Were patients stable in the interim period on the construct to be measured?	Was the time interval appropriate?	Were the test conditions similar for both measurements?	For continuous scores: Was the Standard Error of Measurement (SEM), Smallest Detectable Change (SDC) or Limits of Agreement (LoA) calculated?	Were there any important flaws in the design or methods of the study?
Quek (2002) [86]	Inadequate	Very good	Doubtful	Adequate	Adequate	Inadequate
Rosen (1997) [52]	Adequate	Adequate	Very good	Adequate	Very good	Very good
Utomo (2015) [99]	Adequate	Very good	Very good	Adequate	Very good	Very good

Table D.5. Quality Assessment Known-groups Comparison.

Reference	Overall score (lowest grade)	Was an adequate description provided of important characteristics of the subgroups?	Were design and statistical methods adequate for the hypotheses to be tested?	Were there any other important flaws in the design or statistical methods of the study?
Dargis (2013) [101]	Adequate	Very good	Adequate	Very good
Lim (2003) [102]	Adequate	Very good	Adequate	Very good
Quek (2002) [86]	Inadequate	Very good	Doubtful	Inadequate
Quinta Gomes (2012) [87]	Adequate	Very good	Adequate	Very good
Rosen (1997) [52]	Adequate	Very good	Adequate	Very good
Rosen (1999) [53]	Doubtful	Very good	Adequate	Doubtful
Wiltink (2003) [94]	Adequate	Very good	Adequate	Very good

Table D.6. Quality Assessment Convergent Validity.

Reference	Overall score (lowest grade)	Is it clear what the comparator instrument(s) measure(s)?	Were the measurement properties of the comparator instrument(s) adequate?	Were design and statistical methods adequate for the hypotheses to be tested?	Were there any other important flaws in the design or statistical methods of the study?
Aslan (2011) [95]	Adequate	Very good	Very good	Adequate	Adequate
Cappelleri (2000) [70]	Doubtful	Very good	Doubtful	Adequate	Very good
Cappelleri (2001) [100]	Doubtful	Very good	Doubtful	Adequate	Very good
Cappelleri (2009) [71]	Doubtful	Very good	Very good	Adequate	Doubtful
Garcia-Cruz (2011) [74]	Doubtful	Very good	Very good	Doubtful	Very good

Reference	Overall score (lowest grade)	Is it clear what the comparator instrument(s) measure(s)?	Were the measurement properties of the comparator instrument(s) adequate?	Were design and statistical methods adequate for the hypotheses to be tested?	Were there any other important flaws in the design or statistical methods of the study?
Hwang (2010) [77]	Adequate	Very good	Very good	Adequate	Very good
Mulhall (2008) [80]	Adequate	Very good	Very good	Adequate	Very good
Rosen (1997) [52]	Adequate	Very good	Adequate	Adequate	Very good
Rubio-Aurioles (2009) [89]	Adequate	Very good	Very good	Adequate	Very good
Saffari (2016) [90]	Adequate	Very good	Very good	Adequate	Very good
Wiltink (2003) [94]	Adequate	Very good	Adequate	Adequate	Very good
Flynn (2013) [73]	Adequate	Very good	Very good	Adequate	Very good
Parisot (2014) [84]	Doubtful	Very good	Very good	Adequate	Doubtful
Nimbi (2018) [81]	Doubtful	Very good	Very good	Adequate	Doubtful
Gelhorn (2017) [75]	Doubtful	Very good	Very good	Adequate	Doubtful
Maasoumi (2017) [79]	Adequate	Very good	Very good	Adequate	Very good
O'Toole (2018) [83]	Adequate	Very good	Very good	Adequate	Very good
Pascoal (2017) [85]	Adequate	Very good	Very good	Adequate	Very good
Tang (2018) [92]	Adequate	Very good	Very good	Adequate	Very good

Table D.7. Quality Assessment Divergent Validity.

Reference	Overall score (lowest grade)	Is it clear what the comparator instrument(s) measure(s)?	Were the measurement properties of the comparator instrument(s) adequate?	Were design and statistical methods adequate for the hypotheses to be tested?	Were there any other important flaws in the design or statistical methods of the study?
Dargis (2013) [101]	Adequate	Very good	Adequate	Adequate	Very good
Rosen (1997) [52]	Doubtful	Very good	Doubtful	Adequate	Very good
Wiltink (2003) [94]	Adequate	Very good	Adequate	Adequate	Adequate

Table D.8. Quality Assessment Criterion Validity.

Reference	Overall score (lowest grade)	Can the criterion used or employed be considered as a reasonable 'gold standard'?	for continuous scores: Were correlations, or the area under the receiver operating curve calculated?	for dichotomous scores: Were sensitivity and specificity determined?	Was the sample size included in the analysis to determine the for area under the Receiver Operator Curve (ROC) or sensitivity and specificity adequate?	Were there any important flaws in the design or methods of the study?
Cappelleri (1999) [69]	Very good	Very good	n/a	Very good	Very good	Very good
Lim (2003) [102]	Adequate	Very good	n/a	Very good	Adequate	Very good
Rosen (1999) [53]	Doubtful	Very good	Very good	Very good	Very good	Doubtful
Rubio-Aurioles (2009) [89]	Very good	Very good	n/a	Very good	Very good	Very good
Tang (2015) [98]	Very good	Very good	n/a	Very good	Very good	Very good
Wiltink (2003) [94]	Adequate	Very good	n/a	Very good	Adequate	Very good
Terrier (2017) [93]	Doubtful	Doubtful	n/a	Very good	Very good	Very good

Table D.9. Quality Assessment Responsiveness (construct approach 3).

Reference	Overall score (lowest grade)	For construct approach 3: Was an adequate description provided of the intervention given?	For construct approach 3: Were design and statistical methods adequate for the hypotheses to be tested?	For construct approach 3: Were there any other important flaws in the design or statistical methods of the study?
Althof (2006) [65]	Adequate	Very good	Adequate	Very good
O'Leary (2006) [82]	Adequate	Very good	Adequate	Very good
Quek (2002) [86]	Inadequate	Very good	Doubtful	Inadequate
Rosen (2011) [88]	Very good	n/a	n/a	n/a
Rosen (1997) [52]	Adequate	Very good	Adequate	Very good
Utomo (2015) [99]	Doubtful	Doubtful	Doubtful	Very good
Parisot (2014) [84]	Adequate	Very good	Adequate	Adequate

Table D.10. Quality Assessment Responsiveness (construct approach 2).

Reference	Overall score (lowest grade)	For construct approach 2: Was an adequate description provided of important characteristics of the subgroups?	For construct approach 2: Were design and statistical methods adequate for the hypotheses to be tested?	For construct approach 2: Were there any other important flaws in the design or statistical methods of the study?
Cappelleri (2000) [70]	Adequate	Very good	Adequate	Very good
Cappelleri (2001) [100]	Adequate	Very good	Adequate	Adequate

Table D.11. Quality Assessment Responsiveness (criterion approach).

Reference [88]	Overall score (lowest grade)	For criterion approach: Can the criterion for change be considered as a reasonable gold standard?	For criterion approach: For continuous scores: Were correlations between change scores, or the area under the Receiver Operator Curve (ROC) curve calculated?	For criterion approach: Was the sample size included in the analysis to determine the for area under the Receiver Operator Curve (ROC) or sensitivity and specificity adequate?	For criterion approach: Were there any other important flaws in the design or statistical methods of the study?
Rosen (2011)	Very good	Very good	Very good	Very good	Very good

Appendix E: Methodological Quality Assessments FSFI

Table E.1. Quality Assessment Structural Validity.

Reference	Overall score (lowest grade)	Does the scale consist of effect indicators, i.e. is it based on a reflective model?	for CTT: Was exploratory or confirmatory factor analysis performed?	for IRT: does the chosen model fit to the research question?	Was the sample size included in the analysis adequate?	Were there any important flaws in the design or methods of the study?
Anis (2011) [111]	Adequate	Yes	Adequate	n/a	Very good	Very good
Bartula (2015) [114]	Very good	Yes	Very good	n/a	Very good	Very good
Bartula (2015) b [183]	Adequate	Yes	Adequate	n/a	Very good	Very good
Baser (2012) [115]	Adequate	Yes	Adequate	n/a	Very good	Very good
Burri (2010) [184]	Adequate	Yes	Very good	n/a	Very good	Adequate
Burri (2018) [119]	Adequate	Yes	Adequate	n/a	Very good	Very good
Carpenter (2016) [121]	Adequate	Yes	n/a	Very good	Adequate	Very good
Chang (2009) [122]	Adequate	Yes	Adequate	n/a	Adequate	Very good
Fakhri (2012) [128]	Adequate	Yes	Very good	n/a	Very good	Adequate
Forbes (2014) [104]	Adequate	Yes	Adequate	n/a	Very good	Very good
Ghassamia (2013) [132]	Adequate	Yes	Adequate	n/a	Very good	Very good
Heng (2013) [134]	Adequate	Yes	Adequate	n/a	Very good	Very good
Hevesi (2017) [137]	Very good	Yes	Very good	n/a	Very good	Very good
Ismail (2014) [138]	Adequate	Yes	Adequate	n/a	Very good	Very good
Kalmbach (2015) [140]	Very good	Yes	Very good	n/a	Very good	Very good
Liu (2016) [143]	Adequate	Yes	Very good	n/a	Very good	Adequate
Nowosielski (2013) [150]	Adequate	Yes	Adequate	n/a	Very good	Very good
Opperman (2013) [151]	Inadequate	Yes	Very good	n/a	Inadequate	Very good
Rehman (2015) [154]	Adequate	Yes	Adequate	n/a	Adequate	Very good
Rosen (2000) a [54]	Very good	Yes	Very good	n/a	Very good	Very good

Reference	Overall score (lowest grade)	Does the scale consist of effect indicators, i.e. is it based on a reflective model?	for CTT: Was exploratory or confirmatory factor analysis performed?	for IRT: does the chosen model fit to the research question?	Was the sample size included in the analysis adequate?	Were there any important flaws in the design or methods of the study?
Rosen (2000) b [54]	Inadequate	Yes	Inadequate	n/a	Inadequate	Very good
Rillon-Tabil (2013) [156]	Inadequate	Yes	Adequate	n/a	Inadequate	Very good
Sun (2011) [164]	Adequate	Yes	Adequate	n/a	Very good	Very good
Takahashi (2011) [165]	Adequate	Yes	Adequate	n/a	Adequate	Very good
Ter Kuile (2006) [166]	Adequate	Yes	Adequate	n/a	Very good	Very good
Vallejo-Medina (2018) [169]	Adequate	Yes	Very good	n/a	Very good	Adequate
Wiegel (2005) [173]	Adequate	Yes	Adequate	n/a	Very good	Very good
Witting (2008) [174]	Adequate	Yes	Very good	n/a	Very good	Adequate
Wolpe (2017) [175]	Adequate	Yes	Adequate	n/a	Very good	Very good
Wylomanski (2014) [176]	Adequate	Yes	Very good	n/a	Very good	Adequate

Table E.2. Quality Assessment Internal Consistency.

Reference	Overall score (lowest grade)	Does the scale consist of effect indicators, i.e. is it based on a reflective model?	Was an internal consistency statistic calculated for each unidimensional (sub) scale separately?	For continuous scores: Was Cronbach's alpha or omega calculated?	Were there any important flaws in the design or methods of the study?
Achimas-Cadariu (2013) [109]	Very good	Yes	Very good	Very good	Very good
Anis (2011) [111]	Very good	Yes	Very good	Very good	Very good
Bartula (2015) [114]	Very good	Yes	Very good	Very good	Very good
Bartula (2015)b [183]	Very good	Yes	Very good	Very good	Very good
Baser (2012) [115]	Very good	Yes	Very good	Very good	Very good
Burri (2010) [184]	Very good	Yes	Very good	Very good	Very good
Carvalho (2012) [185]	Very good	Yes	Very good	Very good	Very good
Chang (2009) [122]	Inadequate	Yes	Inadequate	Very good	Very good
Chedraui (2012) [179]	Very good	Yes	Very good	Very good	Very good
Clayton (2010) [124]	Doubtful	Yes	Doubtful	Very good	Very good
Fakhri (2012) [128]	Very good	Yes	Very good	Very good	Very good

Reference	Overall score (lowest grade)	Does the scale consist of effect indicators, i.e. is it based on a reflective model?	Was an internal consistency statistic calculated for each unidimensional (sub) scale separately?	For continuous scores: Was Cronbach's alpha or omega calculated?	Were there any important flaws in the design or methods of the study?
Filocamo (2014) [131]	Very good	Yes	Very good	Very good	Very good
Forbes (2014) [104]	Very good	Yes	Very good	Very good	Very good
Gerstenberger (2010) [133]	Very good	Yes	Very good	Very good	Very good
Ghassamia (2013) [132]	Doubtful	Yes	Doubtful	Very good	Very good
Hevesi (2017) [137]	Very good	Yes	Very good	Very good	Very good
Isidori (2010) [55]	Very good	Yes	Very good	Very good	Very good
Kalmbach (2015) [140]	Very good	Yes	Very good	Very good	Very good
Lee (2014) [180]	Inadequate	Yes	Inadequate	Very good	Very good
Likes (2006) [141]	Adequate	Yes	Very good	Very good	Adequate
Liu (2016) [143]	Very good	Yes	Very good	Very good	Very good
Meston (2003) [145]	Very good	Yes	Very good	Very good	Very good
Nowosielski (2013) [150]	Very good	Yes	Very good	Very good	Very good
Opperman (2013) [151]	Adequate	Yes	Very good	Very good	Adequate
Perez-Lopez (2012) [182]	Very good	Yes	Very good	Very good	Very good
Rehman (2015) [154]	Very good	Yes	Very good	Very good	Very good
Rillon-Tabil (2013) [156]	Adequate	Yes	Very good	Very good	Adequate
Rosen (2000) [54]	Very good	Yes	Very good	Very good	Very good
Ryding (2015) [159]	Very good	Yes	Very good	Very good	Very good
Sidi (2007) [161]	Very good	Yes	Very good	Very good	Very good
Stephenson (2016) [163]	Adequate	Yes	Very good	Very good	Adequate
Sun (2011) [164]	Very good	Yes	Very good	Very good	Very good
Takahashi (2011) [165]	Very good	Yes	Very good	Very good	Very good
Ter Kuile (2006) [166]	Very good	Yes	Very good	Very good	Very good
Trudel (2012) [167]	Very good	Yes	Very good	Very good	Very good
Vallejo-Medina (2018) [169]	Very good	Yes	Very good	Very good	Very good
Verit (2007) [171]	Very good	Yes	Very good	Very good	Very good
Wiegel (2005) [173]	Very good	Yes	Very good	Very good	Very good

Reference	Overall score (lowest grade)	Does the scale consist of effect indicators, i.e. is it based on a reflective model?	Was an internal consistency statistic calculated for each unidimensional (sub) scale separately?	For continuous scores: Was Cronbach's alpha or omega calculated?	Were there any important flaws in the design or methods of the study?
Witting (2008) [174]	Very good	Yes	Very good	Very good	Very good
Wylomanski (2014) [176]	Very good	Yes	Very good	Very good	Very good
Zachariou (2017) [177]	Inadequate	Yes	Inadequate	Very good	Very good

Table E.3. Quality Assessment Reliability.

Reference	Overall score (lowest grade)	Were patients stable in the interim period on the construct to be measured?	Was the time interval appropriate?	Were the test conditions similar for both measurements?	for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	Were there any important flaws in the design or methods of the study?
Anis (2011) [111]	Doubtful	Adequate	Very good	Adequate	Doubtful	Adequate
Bartula (2015) [114]	Doubtful	Adequate	Very good	Adequate	Doubtful	Very good
Borello-France (2008) [117]	Adequate	Adequate	Very good	Adequate	Adequate	Very good
Chang (2009) [122]	Doubtful	Adequate	Very good	Adequate	Doubtful	Adequate
Fakhri (2012) [128]	Adequate	Adequate	Very good	Adequate	Adequate	Very good
Filocamo (2014) [131]	Doubtful	Adequate	Very good	Very good	Doubtful	Very good
Ghassamia (2013) [132]	Doubtful	Adequate	Very good	Adequate	Doubtful	Adequate
Isidori (2010) [55]	Doubtful	Adequate	Very good	Adequate	Doubtful	Very good
Lee (2014) [180]	Adequate	Adequate	Very good	Adequate	Adequate	Very good
Liu (2016) [143]	Doubtful	Adequate	Very good	Adequate	Doubtful	Very good
Nowosielski (2013) [150]	Adequate	Adequate	Very good	Adequate	Adequate	Very good
Rehman (2015) [154]	Doubtful	Adequate	Very good	Adequate	Adequate	Doubtful
Rillon-Tabil (2013) [156]	Doubtful	Adequate	Very good	Adequate	Doubtful	Very good
Rosen (2000) [54]	Doubtful	Adequate	Very good	Adequate	Doubtful	Very good
Ryding (2015) [159]	Doubtful	Adequate	Very good	Adequate	Doubtful	Very good
Sidi (2007) [161]	Doubtful	Adequate	Very good	Very good	Doubtful	Adequate

Reference	Overall score (lowest grade)	Were patients stable in the interim period on the construct to be measured?	Was the time interval appropriate?	Were the test conditions similar for both measurements?	for continuous scores: Was an intraclass correlation coefficient (ICC) calculated?	Were there any important flaws in the design or methods of the study?
Sun (2011) [164]	Doubtful	Adequate	Very good	Adequate	Doubtful	Very good
Takahashi (2011) [165]	Inadequate	Adequate	Very good	Adequate	Adequate	Inadequate
Ter Kuile (2006) [166]	Inadequate	Adequate	Very good	Inadequate	Doubtful	Very good
Verit (2007) [171]	Doubtful	Adequate	Very good	Adequate	Doubtful	Very good
Wolpe (2017) [175]	Adequate	Adequate	Very good	Adequate	Adequate	Very good
Wylomanski (2014) [176]	Adequate	Adequate	Very good	Adequate	Adequate	Very good
Zachariou (2017) [177]	Doubtful	Adequate	Very good	Adequate	Doubtful	Very good

Table E.4. Quality Assessment Known-groups Comparison.

Reference	Overall score (lowest grade)	Was an adequate description provided of important characteristics of the subgroups?	Were design and statistical methods adequate for the hypotheses to be tested?	Were there any other important flaws in the design or statistical methods of the study?
Achimas-Cadariu (2013) [109]	Adequate	Very good	Adequate	Very good
Anis (2011) [111]	Adequate	Adequate	Adequate	Very good
Baser (2012) [115]	Adequate	Adequate	Adequate	Very good
Clayton (2010) [124]	Adequate	Very good	Adequate	Very good
Fakhri (2012) [128]	Adequate	Very good	Adequate	Very good
Ghassamia (2013) [132]	Adequate	Adequate	Adequate	Very good
Likes (2006) [141]	Adequate	Very good	Adequate	Adequate
Meston (2005) [146]	Adequate	Very good	Adequate	Very good
Meston (2003) [145]	Adequate	Very good	Adequate	Very good
Nowosielski (2013) [150]	Adequate	Very good	Adequate	Very good
Rellini (2006) [155]	Adequate	Adequate	Adequate	Doubtful
Rillon-Tabil (2013) [156]	Adequate	Very good	Adequate	Adequate
Rosen (2000) [54]	Adequate	Very good	Adequate	Very good
Ryding (2015) [159]	Adequate	Very good	Adequate	Very good
Sidi (2007) [161]	Adequate	Very good	Adequate	Very good
Sun (2011) [164]	Adequate	Very good	Adequate	Very good

Reference	Overall score (lowest grade)	Was an adequate description provided of important characteristics of the subgroups?	Were design and statistical methods adequate for the hypotheses to be tested?	Were there any other important flaws in the design or statistical methods of the study?
Takahashi (2011) [165]	Adequate	Adequate	Adequate	Very good
Ter Kuile (2006) [166]	Adequate	Very good	Adequate	Very good
Trudel (2012) [167]	Adequate	Adequate	Adequate	Very good
Verit (2007) [171]	Adequate	Very good	Adequate	Very good
Wiegel (2005) [173]	Adequate	Adequate	Adequate	Very good
Wylomanski (2014) [176]	Adequate	Very good	Adequate	Very good
Zachariou (2017) [177]	Adequate	Very good	Adequate	Very good

Table E.5. Quality Assessment Convergent Validity.

Reference	Overall score (lowest grade)	Is it clear what the comparator instrument(s) measure(s)?	Were the measurement properties of the comparator instrument(s) adequate?	Were design and statistical methods adequate for the hypotheses to be tested?	Were there any other important flaws in the design or statistical methods of the study?
Ahmed (2017) [110]	Adequate	Very good	Very good	Adequate	Very good
Aydin (2016) [112]	Adequate	Very good	Very good	Adequate	Very good
Azimi Nekoo (2014) [113]	Adequate	Very good	Very good	Adequate	Very good
Bartula (2015) [114]	Adequate	Very good	Very good	Adequate	Very good
Bartula (2015)b [183]	Doubtful	Doubtful	Doubtful	Adequate	Very good
Baser (2012) [115]	Adequate	Very good	Very good	Adequate	Very good
Bloemendaal (2015) [116]	Adequate	Very good	Very good	Adequate	Very good
Bornefeld-Ettmann (2018) [118]	Adequate	Very good	Adequate	Adequate	Very good
Burri (2010) [184]	Adequate	Very good	Adequate	Adequate	Very good
Burri (2018) [119]	Adequate	Very good	Adequate	Adequate	Very good
Carpenter (2015) [120]	Adequate	Very good	Very good	Adequate	Very good
Chedraui (2012) [179]	Doubtful	Very good	Doubtful	Adequate	Very good
Clayton (2006) [123]	Adequate	Very good	Very good	Adequate	Very good
Clayton (2010) [124]	Adequate	Very good	Adequate	Adequate	Very good
Constantine (2017) [125]	Doubtful	Adequate	Doubtful	Doubtful	Very good
DeRogatis (2010) [126]	Adequate	Very good	Adequate	Adequate	Very good
Eaton (2017) [127]	Adequate	Very good	Adequate	Adequate	Very good
Fakhri (2012) [128]	Doubtful	Adequate	Doubtful	Adequate	Very good
Farkas (2016) [129]	Adequate	Very good	Very good	Adequate	Very good

Reference	Overall score (lowest grade)	Is it clear what the comparator instrument(s) measure(s)?	Were the measurement properties of the comparator instrument(s) adequate?	Were design and statistical methods adequate for the hypotheses to be tested?	Were there any other important flaws in the design or statistical methods of the study?
Ferguson (2012) [130]	Adequate	Very good	Very good	Adequate	Very good
Flynn (2013) [73]	Adequate	Very good	Very good	Adequate	Very good
Ghassamia (2013) [132]	Adequate	Very good	Adequate	Adequate	Very good
Herbenick (2011) [136]	Adequate	Very good	Very good	Adequate	Very good
Herbenick (2010) [135]	Adequate	Very good	Very good	Adequate	Very good
Jing (2018) [139]	Adequate	Very good	Adequate	Adequate	Very good
Lee (2014) [180]	Adequate	Very good	Adequate	Adequate	Very good
Liu (2014) [142]	Adequate	Adequate	Doubtful	Adequate	Very good
Meston (2005) [146]	Inadequate	Very good	Inadequate	Adequate	Very good
Meston (2003) [145]	Adequate	Very good	Adequate	Adequate	Very good
Mestre (2017) [147]	Adequate	Very good	Very good	Adequate	Very good
Mitchell (2012) [181]	Adequate	Very good	Very good	Adequate	Very good
Mohammadi (2014) [148]	Adequate	Very good	Very good	Adequate	Very good
Mohammed (2014) [149]	Adequate	Very good	Very good	Adequate	Very good
Nimbi (2018) [81]	Adequate	Very good	Very good	Adequate	Very good
Nowosielski (2013) [150]	Doubtful	Adequate	Doubtful	Adequate	Very good
Pakpour (2014) [153]	Adequate	Very good	Very good	Adequate	Very good
Pakpour (2013) [152]	Adequate	Very good	Very good	Adequate	Very good
Pascoal (2017) [85]	Doubtful	Very good	Adequate	Adequate	Very good
Rellini (2006) [155]	Doubtful	Very good	Doubtful	Adequate	Doubtful
Perez-Lopez (2012) [182]	Adequate	Very good	Very good	Adequate	Very good
Rogers (2013) [157]	Adequate	Very good	Very good	Adequate	Very good
Rosen (2009) [158]	Adequate	Very good	Adequate	Adequate	Very good
Ryding (2015) [159]	Adequate	Very good	Adequate	Adequate	Very good
Selcuk (2016) [160]	Adequate	Very good	Very good	Adequate	Very good
Sills (2005) [162]	Adequate	Very good	Adequate	Adequate	Very good
Stephenson (2016) [163]	Adequate	Very good	Adequate	Adequate	Very good
Takahashi (2011) [165]	Adequate	Adequate	Adequate	Adequate	Very good
Trutnovsky (2016) [168]	Adequate	Very good	Very good	Adequate	Very good
Vallejo-Medina (2018) [169]	Adequate	Very good	Adequate	Adequate	Very good

Reference	Overall score (lowest grade)	Is it clear what the comparator instrument(s) measure(s)?	Were the measurement properties of the comparator instrument(s) adequate?	Were design and statistical methods adequate for the hypotheses to be tested?	Were there any other important flaws in the design or statistical methods of the study?
Velten (2016) [170]	Adequate	Very good	Very good	Adequate	Very good
Wang (2015) [172]	Adequate	Very good	Very good	Adequate	Very good
Witting (2008) [174]	Adequate	Very good	Very good	Adequate	Very good
Zohre (2014) [178]	Adequate	Very good	Very good	Adequate	Very good

Table E.6. Quality Assessment Divergent Validity.

Reference	Overall score (lowest grade)	Is it clear what the comparator instrument(s) measure(s)?	Were the measurement properties of the comparator instrument(s) adequate?	Were design and statistical methods adequate for the hypotheses to be tested?	Were there any other important flaws in the design or statistical methods of the study?
Achimas-Cadariu (2013) [109]	Adequate	Very good	Adequate	Adequate	Very good
Bartula (2015) [114]	Adequate	Very good	Adequate	Adequate	Very good
Bartula (2015)b [183]	Adequate	Very good	Adequate	Adequate	Very good
Likes (2006) [141]	Adequate	Very good	Adequate	Adequate	Adequate
Nowosielski (2013) [150]	Doubtful	Adequate	Doubtful	Adequate	Very good
Rosen (2000) [54]	Doubtful	Very good	Doubtful	Adequate	Very good
Ryding (2015) [159]	Doubtful	Very good	Doubtful	Adequate	Very good
Trudel (2012) [167]	Doubtful	Very good	Doubtful	Adequate	Very good

Table E.7. Quality Assessment Criterion Validity.

Reference	Overall score (lowest grade)	Can the criterion used or employed be considered as a reasonable 'gold standard'?	for continuous scores: Were correlations, or the area under the receiver operating curve calculated?	for dichotomous scores: Were sensitivity and specificity determined?	Was the sample size included in the analysis to determine the for area under the Receiver Operator Curve (ROC) or sensitivity and specificity adequate?	Were there any important flaws in the design or methods of the study?
Anis (2011) [111]	Very good	Very good	Very good	Very good	Very good	Very good
Fakhri (2012) [128]	Very good	Very good	Very good	Very good	Very good	Very good

Reference	Overall score (lowest grade)	Can the criterion used or employed be considered as a reasonable 'gold standard'?	for continuous scores: Were correlations, or the area under the receiver operating curve calculated?	for dichotomous scores: Were sensitivity and specificity determined?	Was the sample size included in the analysis to determine the for area under the Receiver Operator Curve (ROC) or sensitivity and specificity adequate?	Were there any important flaws in the design or methods of the study?
Gerstenberger (2010) [133]	Very good	Very good	Very good	Very good	Very good	Very good
Isidori (2010) [55]	Very good	Very good	Very good	Very good	Very good	Very good
Lee (2014) [180]	Very good	Very good	Very good	Very good	Very good	Very good
Ma (2014) [144]	Very good	Very good	Very good	Very good	Very good	Very good
Nowosielski (2013) [150]	Very good	Very good	Very good	Very good	Very good	Very good
Ryding (2015) [159]	Very good	Very good	Very good	Very good	Very good	Very good
Sidi (2007) [161]	Very good	Very good	Very good	Very good	Very good	Very good
Ter Kuile (2006) [166]	Adequate	Adequate	Very good	Very good	Very good	Very good
Wiegel (2005) [173]	Very good	Very good	Very good	Very good	Very good	Very good
Zachariou (2017) [177]	Doubtful	Very good	Very good	Very good	Doubtful	Very good

Appendix F: Sample copy of eHIQ-NL

De e – Health Impact Vragenlijst

Deel 1

In dit onderdeel wordt gevraagd naar ***uw*** *algemene houding tegenover gezondheidsgerelateerde websites.*

In dit gedeelte kan *'gezondheidsgerelateerde websites'* staan voor websites die feitelijke gezondheidsinformatie bevatten, ervaringsverhalen over gezondheid van anderen, blogs over gezondheid of discussieforums over gezondheid.

Start alstublieft met het beantwoorden van de onderstaande vragen.

	Selecteer het vakje dat op u van toepassing is.				
In hoeverre bent u het wel of niet eens met de volgende uitspraken?	Helemaal mee oneens	Mee oneens	Noch mee eens, noch mee oneens	Mee eens	Helemaal mee eens
1. Het internet is een betrouwbare bron om mij te helpen begrijpen wat een arts mij vertelt.					
2. Het internet kan mensen helpen om te weten hoe het is om te leven met een gezondheidsprobleem.					
3. Het internet kan nuttig zijn om mensen te helpen beslissen of hun symptomen belangrijk genoeg zijn om een arts te raadplegen.					
4. Ik zou het internet gebruiken als ik hulp nodig zou hebben bij het maken van een beslissing over mijn gezondheid (bijvoorbeeld of ik een arts zou moeten raadplegen, medicatie zou moeten innemen of andere typen behandelingen zou moeten zoeken).					
5. Ik zou het internet gebruiken om na te gaan of de arts mij passend advies geeft.					
6. Het internet is een goede manier om andere mensen te vinden die vergelijkbare gezondheidsproblemen ervaren.					
7. Het kan behulpzaam zijn om gezondheids-gerelateerde ervaringen van andere mensen op het internet te zien.					
8. Het internet is nuttig als je niet wilt vertellen aan mensen in je omgeving (bijvoorbeeld uw familie of collega's) hoe je je voelt.					

9. Het kan geruststellend zijn om te weten dat ik op elk moment van de dag of nacht terecht kan op gezondheidsgerelateerde websites.					
10. Het internet is een goede manier om andere mensen te vinden die geconfronteerd zijn met gezondheidsgerelateerde beslissingen waar ik mogelijk ook mee wordt geconfronteerd.					
11. Het bekijken van websites over gezondheid stelt me gerust dat ik niet alleen ben met mijn gezondheidszorgen.					

<u>Volg alstublieft de onderstaande instructies op:</u>

1. Klik op de onderstaande link naar de gezondheidsgerelateerde website. Er zal **een nieuwe pagina in uw browser openen.**

2. Neem 10 tot 15 minuten de tijd om naar de onderdelen van de website te surfen die **uw interesse hebben.**

3. Wanneer u klaar bent met surfen op de website, **keer dan terug naar deze pagina en klik op 'doorgaan'** om de resterende vragen te beantwoorden.

Gezondheidsgerelateerde website:

(Houdt u er rekening mee dat als u niet binnen 30 minuten naar deze vragenlijst terugkeert uw sessie zal verlopen)

Voor vragen gerelateerd aan deze vragenlijst, neem alstublieft contact op met:

Deel 2

In dit onderdeel wordt gevraagd naar ***uw mening*** *over de gezondheids-gerelateerde website die u zojuist heeft bekeken.*

	Selecteer het vakje dat op u van toepassing is.				
Denkend aan de website die u net bekeken heeft, in hoeverre bent u het wel of niet eens met de volgende uitspraken?	Helemaal mee oneens	Mee oneens	Noch mee eens, noch mee oneens	Mee eens	Helemaal mee eens
1. De website moedigt mij aan om acties te ondernemen die gunstig kunnen zijn voor mijn gezondheid.					
2. De website heeft een positieve kijk.					
3. De informatie op de website liet een gevoel van verwarring bij me achter.					
4. De website bevat nuttige tips over hoe het leven beter te maken.					
5. De website biedt een breed scala aan informatie.					
6. De taal op de website maakte het gemakkelijk te begrijpen.					
7. Ik voel me meer geneigd om op mezelf te letten na het bezoeken van de website.					
8. Ik heb iets nieuws geleerd van de website.					
9. Ik kan de informatie op de website gemakkelijk begrijpen.					
10. De website bereidt me voor op wat er mogelijk gaat gebeuren met mijn gezondheid.					
11. De mensen die hebben bijgedragen aan de website begrijpen wat voor mij belangrijk is.					
12. Ik vertrouw de informatie op de website.					
13. Ik zou de website raadplegen als ik een beslissing zou moeten nemen over mijn gezondheid.					
14. Ik heb een gevoel van solidariteit met andere mensen die de website gebruiken.					
15. Ik kan me identificeren met andere mensen die de website gebruiken.					
16. In zijn geheel, vind ik de website geruststellend.					
17. Ik waardeer het advies dat gegeven wordt op de website.					

Voor vragen gerelateerd aan deze vragenlijst, neem alstublieft contact op met:

	Selecteer het vakje dat op u van toepassing is.				
Denkend aan de website die u net bekeken heeft, in hoeverre bent u het wel of niet eens met de volgende uitspraken?	Helemaal mee oneens	Mee oneens	Noch mee eens, noch mee oneens	Mee eens	Helemaal mee eens
18. De website geeft me het vertrouwen dat ik in staat ben om met mijn gezondheid om te gaan.					
19. Ik heb het gevoel veel gemeen te hebben met andere mensen die de website gebruiken.					
20. De website geeft mij het vertrouwen om mijn gezondheidszorgen aan anderen uit te leggen.					
21. De website helpt me om een beter begrip te hebben van mijn persoonlijke gezondheid.					
22. De website moedigt mij aan om een actievere rol te spelen in mijn gezondheidszorg.					
23. De website geeft mij meer vertrouwen om mijn gezondheid te bespreken met mensen in mijn omgeving (bijvoorbeeld mijn familie of collega's).					
24. Foto's en andere afbeeldingen op de website werden passend gebruikt.					
25. Ik vond de afbeeldingen op de website verontrustend.					
26. De website is gemakkelijk te gebruiken.					

Voor vragen gerelateerd aan deze vragenlijst, neem alstublieft contact op met:

References

References

[1] Ibáñez V, Silva J, Cauli O. A survey on sleep assessment methods. PeerJ 2018;6:e4849. https://doi.org/10.7717/peerj.4849.

[2] Dugdale DC, Epstein R, Pantilat SZ. Time and the patient-physician relationship. Journal of General Internal Medicine 1999;14 Suppl 1:S34–40. https://doi.org/10.1046/J.1525-1497.1999.00263.X.

[3] Irving G, Neves AL, Dambha-Miller H, Oishi A, Tagashira H, Verho A, et al. International variations in primary care physician consultation time: A systematic review of 67 countries 2017;7. https://doi.org/10.1136/bmjopen-2017-017902.

[4] Kessel P van, Triemstra M, Boer D de, Plass AM. Meten van uitkomsten van zorg met PROMs. Nederlands Tijdschrift Voor Evidence Based Practice 2016;14:8–10. https://doi.org/10.1007/s12468-016-0014-0.

[5] Nivel. Met PROMs en PREMs naar betere zorg n.d.

[6] Oh H, Rizo C, Enkin M, Jadad A, Powell J, Pagliari C. What is eHealth (3): a systematic review of published definitions. Journal of Medical Internet Research 2005;7:e1. https://doi.org/10.2196/jmir.7.1.e1.

[7] World Health Organization. From innovation to implementation - eHealth in the WHO European region. Geneva, Switzerland: 2016.

[8] Hout A van der, Uden-Kraan CF van, Witte BI, Coupé VMH, Jansen F, Leemans CR, et al. Efficacy, cost-utility and reach of an eHealth self-management application 'Oncokompas' that helps cancer survivors to obtain optimal supportive care: study protocol for a randomised controlled trial. Trials 2017;18:228. https://doi.org/10.1186/s13063-017-1952-1.

[9] Lubberding S, Uden-Kraan CF van, Te Velde EA, Cuijpers P, Leemans CR, Verdonck-de Leeuw IM. Improving access to supportive cancer care through an eHealth application: a qualitative needs assessment among cancer survivors. Journal of Clinical Nursing 2015;24:1367–79. https://doi.org/10.1111/jocn.12753.

[10] Jansen F, Uden-Kraan CF van, Zwieten V van, Witte BI, Verdonck-de Leeuw IM. Cancer survivors' perceived need for supportive care and their attitude towards self-management and eHealth. Supportive Care in Cancer 2015;23:1679–88. https://doi.org/10.1007/s00520-014-2514-7.

[11] Duman-Lubberding S, Uden-Kraan CF van, Jansen F, Witte BI, Velden LA van der, Lacko M, et al. Feasibility of an eHealth application "OncoKompas" to improve personalized survivorship cancer care. Supportive Care in Cancer 2016;24:2163–71. https://doi.org/10.1007/s00520-015-3004-2.

[12] Chen M-L, Lin C-C. Cancer Symptom Clusters: A Validation Study. Journal of Pain and Symptom Management 2007;34:590–9. https://doi.org/10.1016/J.JPAINSYMMAN.2007.01.008.

[13] Sanson-Fisher R, Girgis A, Boyes A, Bonevski B, Burton L, Cook P. The unmet supportive care needs of patients with cancer. Cancer 2000;88:226–37. https://doi.org/10.1002/(SICI)1097-0142(20000101)88:1<226::AID-CNCR30>3.0.CO;2-P.

[14] Aaronson NK, Mattioli V, Minton O, Weis J, Johansen C, Dalton SO, et al. Beyond treatment – Psychosocial and behavioural issues in cancer survivorship research and practice. European Journal of Cancer Supplements 2014;12:54–64. https://doi.org/10.1016/J.EJCSUP.2014.03.005.

[15] Warrington L, Absolom K, Velikova G. Integrated care pathways for cancer survivors - A role for patient-reported outcome measures and health informatics 2015;54:600–8. https://doi.org/10.3109/0284186X.2014.995778.

[16] Bennett AV, Jensen RE, Basch E. Electronic patient-reported outcome systems in oncology clinical practice. CA: A Cancer Journal for Clinicians 2012;62:336–47. https://doi.org/10.3322/caac.21150.

[17] Harrington CB, Hansen JA, Moskowitz M, Todd BL, Feuerstein M. It's not over when it's over: Long-term symptoms in cancer survivors-a systematic review 2010;40:163–81. https://doi.org/10.2190/PM.40.2.c.

[18] Mitchell AJ, Chan M, Bhatti H, Halton M, Grassi L, Johansen C, et al. Prevalence of depression, anxiety, and adjustment disorder in oncological, haematological, and palliative-care settings: A meta-analysis of 94 interview-based studies. The Lancet Oncology 2011;12:160–74. https://doi.org/10.1016/S1470-2045(11)70002-X.

[19] Leeuwen M van, Husson O, Alberti P, Arraras JI, Chinot OL, Costantini A, et al. Understanding the quality of life (QOL) issues in survivors of cancer: towards the development of an EORTC QOL cancer survivorship questionnaire. Health and Quality of Life Outcomes 2018;16:114. https://doi.org/10.1186/s12955-018-0920-0.

[20] Schagen SB, Klein M, Reijneveld JC, Brain E, Deprez S, Joly F, et al. Monitoring and optimising cognitive function in cancer patients: Present knowledge and future directions. European Journal of Cancer, Supplement 2014;12:29–40. https://doi.org/10.1016/j.ejcsup.2014.03.003.

[21] Schover LR, Kaaij M van der, Dorst E van, Creutzberg C, Huyghe E, Kiserud CE. Sexual dysfunction and infertility as late effects of cancer treatment. European Journal of Cancer, Supplement 2014;12:41–53. https://doi.org/10.1016/j.ejcsup.2014.03.004.

[22] Melissant HC, Uden-Kraan CF van, Lissenberg-Witte BI, Verdonck-de Leeuw IM. Body changes after cancer: female cancer patients' perceived social support and their perspective on care. Supportive Care in Cancer 2019. https://doi.org/10.1007/s00520-019-04729-w.

[23] Bruera E, Hui D. Integrating Supportive and Palliative Care in the Trajectory of Cancer: Establishing Goals and Models of Care. Journal of Clinical Oncology 2010;28:4013–7. https://doi.org/10.1200/JCO.2010.29.5618.

[24] Boyes AW, Girgis A, D'Este C, Zucca AC. Prevalence and correlates of cancer survivors' supportive care needs 6 months after diagnosis: A population-based cross-sectional study. BMC Cancer 2012;12. https://doi.org/10.1186/1471-2407-12-150.

[25] Vermeire E, Hearnshaw H, Van Royen P, Denekens J. Patient adherence to treatment: Three decades of research. A comprehensive review 2001;26:331–42. https://doi.org/10.1046/j.1365-2710.2001.00363.x.

[26] Haskard Zolnierek KB, Dimatteo MR. Physician communication and patient adherence to treatment: A meta-analysis. Medical Care 2009;47:826–34. https://doi.org/10.1097/MLR.0b013e31819a5acc.

[27] Kaba R, Sooriakumaran P. The evolution of the doctor-patient relationship. International Journal of Surgery 2007;5:57–65. https://doi.org/10.1016/j.ijsu.2006.01.005.

[28] Ware E, Snyder M, Wright R, Davies A. Defining and measuring patient satisfaction with medical care. Evaluation and Program Planning 1983;6:247–63. https://doi.org/10.1016/0149-7189(83)90005-8.

[29] Oberst MT. Patients' perceptions of care. Cancer 1984;53:2366–75.

[30] Baker R. Development of a questionnaire to assess patients' satisfaction with consultations in general practice. British Journal of General Practice 1990;40:487–90.

[31] Rubin H, Ware H, Nelson E, Meterko M. The Patient Judgments of Hospital Quality (PJHQ) Questionnaire. Medical Care 1990;28:S17–8.

[32] Hargraves JL, Hays RD, Cleary PD. Psychometric properties of the Consumer Assessment of Health Plans Study (CAHPS) 2.0 adult core survey. Health Services Research 2003;38:1509–28. https://doi.org/10.1111/j.1475-6773.2003.00190.x.

[33] Brédart A, Bottomley A, Blazeby JM, Conroy T, Coens C, D'Haese S, et al. An international prospective study of the EORTC cancer in-patient satisfaction with care measure (EORTC IN-PATSAT32). European Journal of Cancer 2005;41:2120–31. https://doi.org/10.1016/j.ejca.2005.04.041.

[34] Greenhalgh T, Russell J. Why Do Evaluations of eHealth Programs Fail? An Alternative Set of Guiding Principles. PLoS Medicine 2010;7:e1000360. https://doi.org/10.1371/journal.pmed.1000360.

[35] Kelly L, Jenkinson C, Ziebland S. Measuring the effects of online health information for patients: Item generation for an e-health impact questionnaire. Patient Education and Counseling 2013;93:433–8. https://doi.org/10.1016/j.pec.2013.03.012.

[36] Kelly L, Ziebland S, Jenkinson C. Measuring the effects of online health information: scale validation for the e-Health Impact Questionnaire. Patient Education and Counseling 2015:6–12. https://doi.org/10.1016/j.pec.2015.06.008.

[37] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. Journal of Clinical Epidemiology 2010;63:737–45. https://doi.org/10.1016/J.JCLINEPI.2010.02.006.

[38] Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, Vet HCW de, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Quality of Life Research 2018;27:1147–57. https://doi.org/10.1007/s11136-018-1798-3.

[39] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. Quality of Life Research 2010;19:539–49. https://doi.org/10.1007/s11136-010-9606-8.

[40] Bennett CM, Wolford GL, Miller MB. The principled control of false positives in neuroimaging. Social Cognitive and Affective Neuroscience 2009;4:417–22. https://doi.org/10.1093/scan/nsp053.

[41] Kreisle WH, Modiano M. Leukopenia. In:. Decision making in medicine, Mosby; 2010, pp. 242–3. https://doi.org/10.1016/B978-0-323-04107-2.50090-9.

[42] Roski J, Bo-Linn GW, Andrews TA. Creating Value In Health Care Through Big Data: Opportunities And Policy Implications. Health Affairs 2014;33:1115–22. https://doi.org/10.1377/hlthaff.2014.0147.

[43] Vodenčarević A, Goes M van der, O'Jay A. Predicting Flare Probability in Rheumatoid Arthritis using Machine Learning Methods. Data 2018:187–92. https://doi.org/10.5220/0006930501870192.

[44] Chen M, Hao Y, Hwang K, Wang L, Wang L. Disease Prediction by Machine Learning over Big Data from Healthcare Communities. IEEE Access 2017;5:8869–79. https://doi.org/10.1109/ACCESS.2017.2694446.

[45] Siuly S, Zhang Y. Medical Big Data: Neurological Diseases Diagnosis Through Medical Data Analysis 2016;1:54–64. https://doi.org/10.1007/s41019-016-0011-3.

[46] Kirkova J, Aktas A, Walsh D, Davis MP. Cancer Symptom Clusters: Clinical and Research Methodology. Journal of Palliative Medicine 2011;14:1149–66. https://doi.org/10.1089/jpm.2010.0507.

[47] Ward Sullivan C, Leutwyler H, Dunn LB, Miaskowski C. A review of the literature on symptom clusters in studies that included oncology patients receiving primary or adjuvant chemotherapy. Journal of Clinical Nursing 2018;27:516–45. https://doi.org/10.1111/jocn.14057.

[48] Miaskowski C, Barsevick A, Berger A, Casagrande R, Grady PA, Jacobsen P, et al. Advancing Symptom Science Through Symptom Cluster Research: Expert Panel Proceedings and Recommendations. Journal of the National Cancer Institute 2017;109:djw253. https://doi.org/10.1093/jnci/djw253.

[49] Cruz JA, Wishart DS. Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Informatics 2006;2:59–78. https://doi.org/10.1177/117693510600200030.

[50] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction 2015;13:8–17. https://doi.org/10.1016/j.csbj.2014.11.005.

[51] Neijenhuijs KI. PROMs in Oncokompas 2.0 - A systematic review of measurement properties. Amsterdam: Vrije Universiteit Amsterdam; 2017.

[52] Rosen RC, Riley A, Wagner G, Osterloh IH, Kirkpatrick J, Mishra A. The international index of erectile function (IIEF): a multidimensional scale for assessment of erectile dysfunction. Urology 1997;49:822–30. https://doi.org/10.1016/S0090-4295(97)00238-0.

[53] Rosen RC, Cappelleri JC, Smith MD, Lipsky J, Peña BM. Development and evaluation of an abridged, 5-item version of the International Index of Erectile Function (IIEF-5) as a diagnostic tool for erectile dysfunction. International Journal of Impotence Research 1999;11:319–26. https://doi.org/10.1038/sj.ijir.3900472.

[54] Rosen R, Brown C, Heiman J, Leiblum S. The Female Sexual Function Index (FSFI): a multidimensional self-report instrument for the assessment of female sexual function. Journal of Sex 2000.

[55] Isidori AM, Pozza C, Esposito K, Giugliano D, Morano S, Vignozzi L, et al. Development and validation of a 6-item version of the female sexual function index (FSFI) as a diagnostic tool for female sexual dysfunction. J Sex Med 2010;7:1139–46. https://doi.org/10.1111/j.1743-6109.2009.01635.x.

[56] Hopwood P, Fletcher I, Lee A, Al Ghazal S. A body image scale for use with cancer patients. European Journal of Cancer 2001;37:189–97. https://doi.org/10.1016/S0959-8049(00)00353-1.

[57] Gujral S, Conroy T, Fleissner C, Sezer O, King PM, Avery KN, et al. Assessing quality of life in patients with colorectal cancer: An update of the EORTC quality of life questionnaire. European Journal of Cancer 2007;43:1564–73. https://doi.org/10.1016/J.EJCA.2007.04.005.

[58] Melissant HC, Verdonck-de Leeuw IM, Lissenberg-Witte BI, Konings IR, Cuijpers P, Van Uden-Kraan CF. 'Oncokompas', a web-based self-management application to support patient activation and optimal supportive care: a feasibility study among breast cancer survivors. Acta Oncologica 2018;57:924–34. https://doi.org/10.1080/0284186X.2018.1438654.

[59] Hout A van der, Neijenhuijs KI, Jansen F, Uden-Kraan CF van, Aaronson NK, Groenvold M, et al. Measuring health-related quality of life in colorectal cancer patients: systematic review of measurement properties of the EORTC QLQ-CR29. Supportive Care in Cancer 2019:1–18. https://doi.org/10.1007/s00520-019-04764-7.

[60] Rosen R, Cappelleri J, Gendrano Iii N. The International Index of Erectile Function (IIEF): a state-of-the-science review. International Journal of Impotence Research 2002;14:226–44. https://doi.org/10.1038/sj.ijir.3900857.

[61] Terwee CB, Jansma EP, Riphagen II, De Vet HCW. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Quality of Life Research 2009;18:1115–23. https://doi.org/10.1007/s11136-009-9528-5.

[62] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. COSMIN checklist manual. Manual. VU University Medical Center, 2012.

[63] Terwee CB, Mokkink LB, Knol DL, Ostelo RWJG, Bouter LM, De Vet HCW. Rating the methodological quality in systematic reviews of studies on measurement properties: A scoring system for the COSMIN checklist. Quality of Life Research 2012;21:651–7. https://doi.org/10.1007/s11136-011-9960-1.

[64] Mokkink LB, Vet HCW de, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. Quality of Life Research 2018;27:1171–9. https://doi.org/10.1007/s11136-017-1765-4.

[65] Althof SE, O'Leary MP, Cappelleri JC, Hvidsten K, Stecher VJ, Glina S, et al. Sildenafil Citrate Improves Self-Esteem, Confidence, and Relationships in Men with Erectile Dysfunction: Results from an International, Multi-Center, Double-Blind, Placebo-Controlled Trial. The Journal of Sexual Medicine 2006;3:521–9. https://doi.org/10.1111/j.1743-6109.2006.00234.x.

[66] Bayraktar Z, Atun AI. Despite some comprehension problems the international index of erectile function is a reliable questionnaire in erectile dysfunction. Urologia Internationalis 2012;88:170–6. https://doi.org/10.1159/000335432.

[67] Bayraktar Z, Atun I. Impact of physician assistance on the reliability of the International Index of Erectile Function. Andrologia 2013;45:73–7. https://doi.org/10.1111/j.1439-0272.2012.01312.x.

[68] Bushmakin AG, Cappelleri JC, Symonds T, Stecher VJ. Further Understanding of the International Index of Erectile Function at 15+ Years: Confirmatory Factor Analysis and Multidimensional Scaling. Therapeutic Innovation & Regulatory Science 2014;48:246–54. https://doi.org/10.1177/2168479013500056.

[69] Cappelleri J, Rosen R, Smith M, Quirk F. Some developments on the international index of erectile function (IIEF). Drug Information Journal 1999;33:179–90.

[70] Cappelleri J, Siegel R, Osterloh I, Urology RR, 2000. Relationship between patient self-assessment of erectile function and the erectile function domain of the international index of erectile function. Urology 2000;56:477–81.

[71] Cappelleri JC, Bushmakin AG, Symonds T, Schnetzler G. Scoring Correspondence in Outcomes Related to Erectile Dysfunction Treatment on a 4-Point Scale (SCORE-4). The Journal of Sexual Medicine 2009;6:809–19. https://doi.org/10.1111/j.1743-6109.2008.01155.x.

[72] Coyne K, Mandalia S, McCullough S, Catalan J, Noestlinger C, Colebunders R, et al. The international index of erectile function: Development of an adapted tool for use in HIV-positive men who have sex with men. Journal of Sexual Medicine 2010;7:769–74. https://doi.org/10.1111/j.1743-6109.2009.01579.x.

[73] Flynn KE, Reeve BB, Lin L, Cyranowski JM, Bruner DW, Weinfurt KP. Construct validity of the PROMIS sexual function and satisfaction measures in patients with cancer. Health Qual Life Outcomes 2013;11:1. https://doi.org/10.1186/1477-7525-11-40.

[74] García-Cruz E, Romero Otero J, Martínez Salamanca JI, Leibar Tamayo A, Rodríguez Antolín A, Astobieta Odriozola A, et al. Linguistic and Psychometric Validation of the Erection Hardness Score to Spanish. Journal of Sexual Medicine 2011;8:470–4. https://doi.org/10.1111/j.1743-6109.2010.02124.x.

[75] Gelhorn HL, Roberts LJ, Khandelwal N, Revicki DA, DeRogatis LR, Dobs A, et al. Psychometric Evaluation of the Hypogonadism Impact of Symptoms Questionnaire Short Form (HIS-Q-SF). Journal of Sexual Medicine 2017;14:1046–58. https://doi.org/10.1016/j.jsxm.2017.05.013.

[76] Gonzáles AI, Sties SW, Wittkopf PG, Mara LS de, Ulbrich AZ, Cardoso FL, et al. Validation of the International Index of Erectile Function (IIFE) for use in Brazil. Arquivos Brasileiros de Cardiologia 2013;101:176–81. https://doi.org/10.5935/abc.20130141.

[77] Hwang TIS, Tsai TE, Lin YIC, Chiang HS, Chang LS. A survey of erectile dysfunction in Taiwan: Use of the erection hardness score and quality of erection questionnaire. Journal of Sexual Medicine 2010;7:174. https://doi.org/10.1111/j.1743-6109.2010.01837.x.

[78] Kriston L, Günzler C, Harms A, Berner M. Confirmatory factor analysis of the German version of the International Index of Erectile Function (IIEF): A comparison of four models. Journal of Sexual Medicine 2008;5:92–9. https://doi.org/10.1111/j.1743-6109.2007.00474.x.

[79] Maasoumi R, Mokarami H, Nazifi M, Stallones L, Taban A, Aval MY, et al. Psychometric Properties of the Persian Translation of the Sexual Quality of Life-Male Questionnaire. American Journal of Men's Health 2017;11:564–72. https://doi.org/10.1177/1557988316629641.

[80] Mulhall JP, King R, Kirby M, Hvidsten K, Symonds T, Bushmakin AG, et al. Evaluating the sexual experience in men: Validation of the sexual experience questionnaire. Journal of Sexual Medicine 2008;5:365–76. https://doi.org/10.1111/j.1743-6109.2007.00694.x.

[81] Nimbi FM, Tripodi F, Simonelli C, Nobre P. Sexual Modes Questionnaire (SMQ): Translation and Psychometric Properties of the Italian Version of the Automatic Thought Scale. Journal of Sexual Medicine 2018;15:410–5. https://doi.org/10.1016/j.jsxm.2018.01.002.

[82] O'Leary MP, Althof SE, Cappelleri JC, Crowley A, Sherman N, Duttagupta S. Self-Esteem, Confidence and Relationship Satisfaction of Men With Erectile Dysfunction Treated With Sildenafil Citrate: A Multicenter, Randomized, Parallel Group, Double-Blind, Placebo Controlled Study in the United States. Journal of Urology 2006;175:1058–62. https://doi.org/10.1016/S0022-5347(05)00418-0.

[83] O'Toole A, Silva PS de, Marc LG, Ulysse CA, Testa MA, Ting A, et al. Sexual Dysfunction in Men With Inflammatory Bowel Disease: A New IBD-Specific Scale. Inflammatory Bowel Diseases 2018;24:310–6. https://doi.org/10.1093/ibd/izx053.

[84] Parisot J, Yiou R, Salomon L, Taille A de la, Lingombet O, Audureau E. Erection hardness score for the evaluation of erectile dysfunction: Further psychometric assessment in patients treated by intracavernous prostaglandins injections after radical prostatectomy. Journal of Sexual Medicine 2014;11:2109–18. https://doi.org/10.1111/jsm.12584.

[85] Pascoal PM, Alvarez M-J, Pereira CR, Nobre P. Development and Initial Validation of the Beliefs About Sexual Functioning Scale: A Gender Invariant Measure. Journal of Sexual Medicine 2017;14:613–23. https://doi.org/10.1016/j.jsxm.2017.01.021.

[86] Quek KF, Low WY, Razack AH, Chua CB, Loh CS. Reliability and validity of the Malay version of the International Index of Erectile Function (IIEF-15) in the Malaysian population. International Journal of Impotence 2002;14:310–5. https://doi.org/10.1038/sj.ijir.3900902.

[87] Quinta Gomes AL, Nobre P. The International Index of Erectile Function (IIEF-15): Psychometric Properties of the Portuguese Version. Journal of Sexual Medicine 2012;9:180–7. https://doi.org/10.1111/j.1743-6109.2011.02467.x.

[88] Rosen RC, Allen KR, Ni X, Araujo AB. Minimal clinically important differences (MCID) in the erectile function (EF) domain of the international index of erectile function (IIEF). Journal of Urology 2011;185:e615. https://doi.org/10.1016/j.juro.2011.02.1527.

[89] Rubio-Aurioles E, Sand M, Terrein-Roccatti N, Dean J, Longworth J, Eardley I, et al. Female assessment of male erectile dysfunction detection scale (FAME): Development and validation. Journal of Sexual Medicine 2009;6:2255–70. https://doi.org/10.1111/j.1743-6109.2009.01348.x.

[90] Saffari M, Pakpour AH, Burri A. Cross-Cultural Adaptation of the Male Genital Self-Image Scale in Iranian Men. Sexual Medicine 2016;4:e34–42. https://doi.org/10.1016/j.esxm.2015.12.005.

[91] Serefoglu EC, Atmaca AF, Dogan B, Altinova S, Akbulut Z, Balbay MD. Problems in understanding the Turkish translation of the international index of Erectile Function. Journal of Andrology 2008;29:369–73. https://doi.org/10.2164/jandrol.107.004366.

[92] Tang D-D, Li C, Peng D-W, Zhang X-S. Validity of premature ejaculation diagnostic tool and its association with International Index of Erectile Function-15 in Chinese men with evidence-based-defined premature ejaculation. Asian Journal of Andrology 2018;20:19–23. https://doi.org/10.4103/aja.aja_9_17.

[93] Terrier JE, Mulhall JP, Nelson CJ. Exploring the Optimal Erectile Function Domain Score Cutoff That Defines Sexual Satisfaction After Radical Prostatectomy. Journal of Sexual Medicine 2017;14:804–9. https://doi.org/10.1016/j.jsxm.2017.04.672.

[94] Wiltink J, Hauck EW, Phädayanon M, Weidner W, Beutel ME. Validation of the German version of the International Index of Erectile Function (IIEF) in patients with erectile dysfunction, Peyronie's disease and controls. International Journal of Impotence Research 2003;15:192–7. https://doi.org/10.1038/sj.ijir.3900997.

[95] Aslan Y, Tuncel A, Aydin O, Balci M, Karabulut E, Atan A. The association between erection hardness grading scale and international index of erectile function in men with erectile dysfunction treated with sildenafil citrate. Urologia Internationalis 2011;86:434–8. https://doi.org/10.1159/000324100.

[96] Lin CY, Pakpour AH, Burri A, Montazeri A. Rasch Analysis of the Premature Ejaculation Diagnostic Tool (PEDT) and the International Index of Erectile Function (IIEF) in an Iranian Sample of Prostate Cancer Patients. PLoS ONE 2016;11:e0157460. https://doi.org/10.1371/journal.pone.0157460.

[97] Mahmood MA, Ur Rehman K, Khan MA, Sultan T. Translation, Cross-Cultural Adaptation, and Psychometric Validation of the 5-Item International Index of Erectile Function (IIEF-5) into Urdu. Journal of Sexual Medicine 2012;9:1883–6. https://doi.org/10.1111/j.1743-6109.2012.02714.x.

[98] Tang Y, Wang Y, Zhu H, Jiang X, Gan Y, Yang J. Bias in Evaluating Erectile Function in Lifelong Premature Ejaculation Patients with the International Index of Erectile Function-5. Journal of Sexual Medicine 2015;12:2061–9. https://doi.org/10.1111/jsm.12988.

[99] Utomo E, Blok BF, Pastoor H, Bangma CH, Korfage IJ. The measurement properties of the five-item International Index of Erectile Function (IIEF-5): A Dutch validation study. Andrology 2015;3:1154–9. https://doi.org/10.1111/andr.12112.

[100] Cappelleri J, Siegel R, Glasser D, …IOC, 2001. Relationship between patient self-assessment of erectile dysfunction and the sexual health inventory for men. Clinical Therapeutics 2001;23.

[101] Dargis L, Trudel G, Cadieux J, Villeneuve L, Préville M, Boyer R. Validation of the International Index of Erectile Function (IIEF) and presentation of norms in older men. Sexologies 2013;22:e20-6. https://doi.org/10.1016/j.sexol.2012.01.001.

[102] Lim TO, Das A, Rampal S, Zaki M, Sahabudin RM, Rohan MJ, et al. Cross-cultural adaptation and validation of the English version of the International Index of Erectile Function (IIEF) for use in Malaysia. International Journal of Impotence Research 2003;15:329–36. https://doi.org/10.1038/sj.ijir.3901009.

[103] Forbes MK. Response to Rosen et al. (2014) "Commentary on 'Critical Flaws in the FSFI and IIEF'". The Journal of Sex Research 2014;51:498–502. https://doi.org/10.1080/00224499.2014.895795.

[104] Forbes MK, Baillie AJ, Schniering CA. Critical Flaws in the Female Sexual Function Index and the International Index of Erectile Function. The Journal of Sex Research 2014;51:485–91. https://doi.org/10.1080/00224499.2013.876607.

[105] Rosen RC, Revicki DA, Sand M. Commentary on "Critical Flaws in the FSFI and IIEF". The Journal of Sex Research 2014;51:492–7. https://doi.org/10.1080/00224499.2014.894491.

[106] Basson R, Berman J, Burnett A, Derogatis L, Fergusen D, Fourcroy J, et al. Report of the international consensus development conference on female sexual dysfunction: definitions and classifications. The Journal of Urology 2000;163:888–93. https://doi.org/10.1016/S0022-5347(05)67828-7.

[107] Locke HJ, Wallace KM. Short Marital-Adjustment and Prediction Tests: Their Reliability and Validity. Marriage and Family Living 1959;21:251. https://doi.org/10.2307/348022.

[108] American Psychiatric Association., American Psychiatric Association. DSM-5 Task Force. Diagnostic and statistical manual of mental disorders : DSM-5. American Psychiatric Association; 2013.

[109] Achimas-Cadariu P, Irimie A, Iancu M, Pop F, Lancrajan L, Lisencu C. IDENTIFICATION AND VALIDATION OF QUALITY OF LIFE MEASURES IN A POPULATION OF WOMEN WITH PREMALIGNANT AND MALIGNANT PATHOLOGY AT CHILDBEARING AGE. Journal of Cognitive and Behavioural Psychotherapies 2013;13:409–20.

[110] Ahmed MR, Shaaban MM, Meky HK. Assessment of sexually related personal distress accompanying premenopausal sexual dysfunction with an Arabic version of the Female Sexual Distress Scale. International Journal of Gynecology & Obstetrics 2017;139:65–70. https://doi.org/10.1002/ijgo.12255.

[111] Anis T, Gheit AS. Arabic translation of Female Sexual Function Index and validation in an Egyptian population. Journal of Sexual Medicine 2011;8:3370–8. https://doi.org/10.1111/j.1743-6109.2011.02471.x.

[112] Aydin S, Onaran OI, Topalan K, Aydin CA, Dansuk R. Development and Validation of Turkish Version of The Female Sexual Distress Scale-Revised. Sexual Medicine 2016;4:E43–50. https://doi.org/10.1016/j.esxm.2015.12.003.

[113] Azimi Nekoo E, Burri A, Ashrafti F, Fridlund B, Koenig HG, Derogatis LR, et al. Psychometric properties of the iranian version of the female sexual distress scale-revised in women. Journal of Sexual Medicine 2014;11:995–1004. https://doi.org/10.1111/jsm.12449.

[114] Bartula I, Sherman KA. The Female Sexual Functioning Index (FSFI): evaluation of acceptability, reliability, and validity in women with breast cancer. Supportive Care in Cancer 2015;23:2633–41. https://doi.org/10.1007/s00520-015-2623-y.

[115] Baser RE, Li YL, Carter J. Psychometric validation of the female sexual function index (FSFI) in cancer survivors. Cancer 2012;118:4606–18. https://doi.org/10.1002/cncr.26739.

[116] Bloemendaal LBA, Laan ETM. The Psychometric Properties of the Sexual Excitation/Sexual Inhibition Inventory for Women (SESII-W) Within a Dutch Population. Journal of Sex Research 2015;52:69–82. https://doi.org/10.1080/00224499.2013.826166.

[117] Borello-France D, Dusi J, O'Leary M, Misplay S, Okonski J, Leng W, et al. Test-retest reliability of the Urge-Urinary Distress Inventory and Female Sexual Function Index in women with multiple sclerosis. Urologic Nursing : Official Journal of the American Urological Association Allied 2008;28:30–5.

[118] Bornefeld-Ettmann P, Steil R, Hoefling V, Wesslau C, Lieberz KA, Rausch S, et al. Validation of the German Version of the Sexual Self-Esteem Inventory for Women and its Application in a Sample of Sexually and Physically Abused Women. Sex Roles 2018;79:109–22. https://doi.org/10.1007/s11199-017-0849-5.

[119] Burri A, Porst H. Preliminary Validation of a German Version of the Sexual Complaints Screener for Women in a Female Population Sample. Sexual Medicine 2018;6:123–30. https://doi.org/10.1016/j.esxm.2018.01.001.

[120] Carpenter JS, Reed SD, Guthrie KA, Larson JC, Newton KM, Lau RJ, et al. Using an FSDS-R Item to Screen for Sexually Related Distress: A MsFLASH Analysis. Sexual Medicine 2015;3:7–13. https://doi.org/10.1002/sm2.53.

[121] Carpenter JS, Jones SMW, Studts CR, Heiman JR, Reed SD, Newton KM, et al. Female Sexual Function Index Short Version: A MsFLASH Item Response Analysis. Archives of Sexual Behaviour 2016;45:1897–905. https://doi.org/10.1007/s10508-016-0804-5.

[122] Chang S-R, Chang T-C, Chen K-H, Lin H-H. Developing and Validating a Taiwan Version of the Female Sexual Function Index for Pregnant Women. Journal of Sexual Medicine 2009;6:1609–16. https://doi.org/10.1111/j.1743-6109.2009.01247.x.

[123] Clayton AH, Segraves RT, Leiblum S, Basson R, Pyke R, Cotton D, et al. Reliability and validity of the Sexual Interest and Desire Inventory-Female (SIDI-F), a scale designed to measure severity of female hypoactive sexual desire disorder. Journal of Sex and Marital Therapy 2006;32:115–35. https://doi.org/10.1080/00926230500442300.

[124] Clayton AH, Goldmeier D, Nappi RE, Wunderlich G, Lewis-D'Agostino DJ, Pyke R. Validation of the sexual interest and desire inventory-female in hypoactive sexual desire disorder. Journal of Sexual Medicine 2010;7:3918–28. https://doi.org/10.1111/j.1743-6109.2010.02016.x.

[125] Constantine ML, Pauls RN, Rogers RR, Rockwood TH. Validation of a single summary score for the Prolapse/Incontinence Sexual Questionnaire-IUGA revised (PISQ-IR). International Urogynecology Journal 2017;28:1901–7. https://doi.org/10.1007/s00192-017-3373-9.

[126] DeRogatis LR, Allgood A, Auerbach P, Eubank D, Greist J, Bharmal M, et al. Validation of a Women's Sexual Interest Diagnostic Interview - Short Form (WSID-SF) and a Daily Log of Sexual Activities (DLSA) in postmenopausal women with hypoactive sexual desire disorder. J Sex Med 2010;7:917–27. https://doi.org/10.1111/j.1743-6109.2009.01528.x.

[127] Eaton AA, Baser RE, Seidel B, Stabile C, Canty JP, Goldfrank DJ, et al. Validation of Clinical Tools for Vaginal and Vulvar Symptom Assessment in Cancer Patients and Survivors. Journal of Sexual Medicine 2017;14:144–51. https://doi.org/10.1016/j.jsxm.2016.11.317.

[128] Fakhri A, Pakpour AH, Burri A, Morshedi H, Zeidi IM. The Female Sexual Function Index: Translation and Validation of an Iranian Version. Journal of Sexual Medicine 2012;9:514–23. https://doi.org/10.1111/j.1743-6109.2011.02553.x.

[129] Farkas B, Tiringer I, Farkas N, Kenyeres B, Nemeth Z. Hungarian language validation of the Pelvic Organ Prolapse/Incontinence Sexual Questionnaire, IUGA-Revised (PISQ-IR). International Urogynecology Journal 2016;27:1831–6. https://doi.org/10.1007/s00192-016-3047-z.

[130] Ferguson SE, Urowitz S, Massey C, Wegener M, Quartey N, Wiljer D, et al. Confirmatory factor analysis of the sexual adjustment and body image scale in women with gynecologic cancer. Cancer 2012;118:3095–104. https://doi.org/10.1002/cncr.26632.

[131] Filocamo MT, Serati M, Marzi VL, Costantini E, Milanesi M, Pietropaolo A, et al. The Female Sexual Function Index (FSFI): Linguistic Validation of the Italian Version. Journal of Sexual Medicine 2014;11:447–53. https://doi.org/10.1111/jsm.12389.

[132] Ghassamia M, Asghari A, Shaeiri MR, Safarinejad MR. Validation of Psychometric Properties of the Persian Version of the Female Sexual Function Index. Urology Journal 2013;10:878–85.

[133] Gerstenberger EP, Rosen RC, Brewer JV, Meston CM, Brotto LA, Wiegel M, et al. Sexual Desire and the Female Sexual Function Index (FSFI): A Sexual Desire Cutpoint for Clinical Interpretation of the FSFI in Women with and without Hypoactive Sexual Desire Disorder. Journal of Sexual Medicine 2010;7:3096–103. https://doi.org/10.1111/j.1743-6109.2010.01871.x.

[134] Heng YS, Sidi H, Jaafar NRN, Razali R, Ram H. Phases of female sexual response cycle among Malaysian women with Infertility: A factor analysis study. Asia-Pacific Psychiatry 2013;5:50–4. https://doi.org/10.1111/appy.12044.

[135] Herbenick D, Reece M. Development and validation of the female genital self-image scale. Journal of Sexual Medicine 2010;7:1822–30. https://doi.org/10.1111/j.1743-6109.2010.01728.x.

[136] Herbenick D, Schick V, Reece M, Sanders S, Dodge B, Fortenberry JD. The Female Genital Self-Image Scale (FGSIS): Results from a Nationally Representative Probability Sample of Women in the United States. Journal of Sexual Medicine 2011;8:158–66. https://doi.org/10.1111/j.1743-6109.2010.02071.x.

[137] Hevesi K, Meszaros V, Kovi Z, Marki G, Szabo M. Different Characteristics of the Female Sexual Function Index in a Sample of Sexually Active and Inactive Women. Journal of Sexual Medicine 2017;14:1133–41. https://doi.org/10.1016/j.jsxm.2017.07.008.

[138] Ismail AH, Bau R, Sidi H, Guan NC, Naing L, Jaafar NR, et al. Factor analysis study on sexual responses in women with Type 2 diabetes mellitus. Comprehensive Psychiatry 2014;55:S34–7. https://doi.org/10.1016/j.comppsych.2012.12.028.

[139] Jing L-w, Zhang C, Jin F, Wang A-p. Development of a Quality of Sexual Life Questionnaire for Breast Cancer Survivors in Mainland China. Medical Science Monitor 2018;24:4101–12. https://doi.org/10.12659/MSM.906666.

[140] Kalmbach DA, Ciesla JA, Janata JW, Kingsberg SA. The Validation of the Female Sexual Function Index, Male Sexual Function Index, and Profile of Female Sexual Function for Use in Healthy Young Adults. Archives of Sexual Behavior 2015;44:1651–62. https://doi.org/10.1007/s10508-014-0334-y.

[141] Likes WM, Stegbauer C, Hathaway D, Brown C, Tillmanns T. Use of the female sexual function index in women with vulvar intraepithelial neoplasia. Journal of Sex & Marital Therapy 2006;32:255–66. https://doi.org/10.1080/00926230600575348.

[142] Liu B, Su M, Zhan H, Yang F, Li W, Zhou X. Adding a sexual dysfunction domain to UPOINT system improves association with symptoms in women with interstitial cystitis and bladder pain syndrome. Urology 2014;84:1308–13. https://doi.org/10.1016/j.urology.2014.08.018.

[143] Liu H, Yu J, Chen Y, He P, Zhou L, Tang X, et al. Sexual function in cervical cancer patients: Psychometric properties and performance of a Chinese version of the Female Sexual Function Index. European Journal of Oncology Nursing : The Official Journal of European Oncology Nursing Society 2016;20:24–30. https://doi.org/10.1016/j.ejon.2015.06.007.

[144] Ma J, Pan L, Lei Y, Zhang A, Kan Y. Prevalence of female sexual dysfunction in urban chinese women based on cutoff scores of the chinese version of the female sexual function index: A preliminary study. Journal of Sexual Medicine 2014;11:909–19. https://doi.org/10.1111/jsm.12451.

[145] Meston CM. Validation of the female sexual function index (FSFI) in women with female orgasmic disorder and in women with hypoactive sexual desire disorder. Journal of Sex and Marital Therapy 2003;29:39–46. https://doi.org/10.1080/713847100.

[146] Meston C, Trapnell P. Development and validation of a five-factor sexual satisfaction and distress scale for women: The Sexual Satisfaction Scale for Women (SSS-W). Journal of Sexual Medicine 2005;2:66–81. https://doi.org/10.1111/j.1743-6109.2005.20107.x.

[147] Mestre M, Lleberia J, Pubill J, Espuña-Pons M. Spanish version of the Pelvic Organ Prolapse/ Urinary Incontinence Sexual Questionnaire IUGA-Revised (PISQ-IR): Transcultural validation. International Urogynecology Journal 2017;28:1865–73. https://doi.org/10.1007/s00192-017-3312-9.

[148] Mohammadi K, Rahnama P, Montazeri A, Foley FW. The multiple sclerosis intimacy and sexuality questionnaire-19: Reliability, validity, and factor structure of the persian version. Journal of Sexual Medicine 2014;11:2225–31. https://doi.org/10.1111/jsm.12531.

[149] Mohammed GFE-K, Hassan H. Validity and reliability of the arabic version of the female genital self-image scale. Journal of Sexual Medicine 2014;11:1193–200. https://doi.org/10.1111/jsm.12494.

[150] Nowosielski K, Wróbel B, Sioma-Markowska U, Poreba R. Development and Validation of the Polish Version of the Female Sexual Function Index in the Polish Population of Females. Journal of Sexual Medicine 2013;10:386–95. https://doi.org/10.1111/jsm.12012.

[151] Opperman EA, Benson LE, Milhausen RR. Confirmatory Factor Analysis of the Female Sexual Function Index. Journal of Sex Research 2013;50:29–36. https://doi.org/10.1080/00224499.2011.628423.

[152] Pakpour AH, Zeidi IM, Saffari M, Burri A. Psychometric Properties of the Iranian Version of the Sexual Quality of Life Scale among Women. Journal of Sexual Medicine 2013;10:981–9. https://doi.org/10.1111/jsm.12042.

[153] Pakpour AH, Zeidi IM, Ziaeiha M, Burri A. Cross-cultural adaptation of the Female Genital Self-Image Scale (FGSIS) in Iranian female college students. Journal of Sex Research 2014;51:646–53. https://doi.org/10.1080/00224499.2013.821441.

[154] Rehman KU, Asif Mahmood M, Sheikh SS, Sultan T, Khan MA. The Female Sexual Function Index (FSFI): Translation, Validation, and Cross-Cultural Adaptation of an Urdu Version "FSFI-U". Sexual Medicine 2015;3:244–50. https://doi.org/10.1002/sm2.77.

[155] Rellini A, Meston C. The sensitivity of event logs, self-administered questionnaires and photoplethysmography to detect treatment-induced changes in female sexual arousal disorder (FSAD) diagnosis. Journal of Sexual Medicine 2006;3:283–91. https://doi.org/10.1111/j.1743-6109.2005.00153.x.

[156] Rillon-Tabil N, Malong CL, Vicera JJ, Gomez MH. Translation and validity of the female sexual function index Filipino version (FSFI-Fil). Phillippine Journal of Internal Medicine 2013;51.

[157] Rogers RG, Rockwood TH, Constantine ML, Thakar R, Kammerer-Doak DN, Pauls RN, et al. A new measure of sexual function in women with pelvic floor disorders (PFD): The Pelvic Organ Prolapse/Incontinence Sexual Questionnaire, IUGA-Revised (PISQ-IR). International Urogynecology Journal 2013;24:1091–103. https://doi.org/10.1007/s00192-012-2020-8.

[158] Rosen RC, Bachmann GA, Reese JB, Gentner L, Leiblum S, Wajszczuk C, et al. Female sexual well-being scale™ (FSWB Scale™): Development and psychometric validation in sexually functional women. Journal of Sexual Medicine 2009;6:1297–305. https://doi.org/10.1111/j.1743-6109.2009.01240.x.

[159] Ryding EL, Blom C. Validation of the Swedish Version of the Female Sexual Function Index (FSFI) in Women with Hypoactive Sexual Desire Disorder. Journal of Sexual Medicine 2015;12:341–9. https://doi.org/10.1111/jsm.12778.

[160] Selcuk S, Kucukbas M, Cam C, Eser A, Devranoglu B, Turkyilmaz S, et al. Validation of the Turkish Version of the Sexual Health Outcomes in Women Questionnaire (SHOW-Q) in Turkish-Speaking Women. Sexual Medicine 2016;4:e89–94. https://doi.org/10.1016/j.esxm.2016.01.003.

[161] Sidi H, Abdullah N, Puteh SE, Midin M. The female sexual function index (FSFI): Validation of the malay version. Journal of Sexual Medicine 2007;4:1642–54. https://doi.org/10.1111/j.1743-6109.2007.00476.x.

[162] Sills T, Wunderlich G, Pyke R, Segraves RT, Leiblum S, Clayton A, et al. The Sexual Interest and Desire Inventory-Female (SIDI-F): Item response analyses of data from women diagnosed with hypoactive sexual desire disorder. Journal of Sexual Medicine 2005;2:801–18. https://doi.org/10.1111/j.1743-6109.2005.00146.x.

[163] Stephenson KR, Toorabally N, Lyons L, Meston C. Further Validation of the Female Sexual Function Index: Specificity and Associations With Clinical Interview Data. Journal of Sex & Marital Therapy 2016;42:448–61. https://doi.org/10.1080/0092623X.2015.1061078.

[164] Sun X, Li C, Jin L, Fan Y, Wang D. Development and Validation of Chinese Version of Female Sexual Function Index in a Chinese Population-A Pilot Study. Journal of Sexual Medicine 2011;8:1101–11. https://doi.org/10.1111/j.1743-6109.2010.02171.x.

[165] Takahashi M, Inokuchi T, Watanabe C, Saito T, Kai I. The female sexual function index (FSFI): Development of a japanese version. Journal of Sexual Medicine 2011;8:2246–54. https://doi.org/10.1111/j.1743-6109.2011.02267.x.

[166] Ter Kuile M, Brauer M, Laan E. The Female Sexual Function Index (FSFI) and the Female Sexual Distress Scale (FSDS): Psychometric properties within a Dutch population. Journal of Sex and Marital Therapy 2006;32:289–304. https://doi.org/10.1080/00926230600666261.

[167] Trudel G, Dargis L, Cadieux J, Villeneuve L, Préville M, Boyer R. Validation of the Female Sexual Function Index (FSFI) and presentation of norms in older women. Sexologies 2012;21:161–7. https://doi.org/10.1016/j.sexol.2012.01.003.

[168] Trutnovsky G, Nagele E, Ulrich D, Aigmüller T, Dörfler D, Geiss I, et al. German translation and validation of the Pelvic Organ Prolapse/Incontinence Sexual Questionnaire–IUGA revised (PISQ-IR). International Urogynecology Journal 2016;27:1235–44. https://doi.org/10.1007/s00192-016-2969-9.

[169] Vallejo-Medina P, Perez-Duran C, Saavedra-Roa A. Translation, Adaptation, and Preliminary Validation of the Female Sexual Function Index into Spanish (Colombia). Archives of Sexual Behavior 2018;47:797–810. https://doi.org/10.1007/s10508-017-0976-7.

[170] Velten J, Scholten S, Graham CA, Margraf J. Psychometric Properties of the Sexual Excitation/Sexual Inhibition Inventory for Women in a German Sample. Archives of Sexual Behavior 2016;45:303–14. https://doi.org/10.1007/s10508-015-0547-8.

[171] Verit FF, Verit A. Validation of the female sexual function index in women with chronic pelvic pain. Journal of Sexual Medicine 2007;4:1635–41. https://doi.org/10.1111/j.1743-6109.2007.00604.x.

[172] Wang H, Lau H-H, Hung M-J, Huang W-C, Zheng Y-W, Su T-H. Validation of a Mandarin Chinese version of the pelvic organ prolapse/urinary incontinence sexual questionnaire IUGA–revised (PISQ-IR). International Urogynecology Journal and Pelvic Floor Dysfunction 2015;26:1695–700. https://doi.org/10.1007/s00192-015-2744-3.

[173] Wiegel M, Meston C, Rosen R. The Female Sexual Function Index (FSFI): Cross-validation and development of clinical cutoff scores. Journal of Sex and Marital Therapy 2005;31:1–20. https://doi.org/10.1080/00926230590475206.

[174] Witting K, Santtila P, Jern P, Varjonen M, Wager I, Höglund M, et al. Evaluation of the Female Sexual Function Index in a population based sample from Finland. Archives of Sexual Behavior 2008;37:912–24. https://doi.org/10.1007/s10508-007-9287-8.

[175] Wolpe RE, Queiroz AP, Zomkowski K, Sperandio FF. Psychometric properties of the Female Sexual Function Index in the visual analogue scale format. Sexual Health 2017;14:213–20. https://doi.org/10.1071/SH16131.

[176] Wylomanski S, Bouquin R, Philippe H-J, Poulin Y, Hanf M, Dréno B, et al. Psychometric properties of the French Female Sexual Function Index (FSFI). Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation 2014;23:2079–87. https://doi.org/10.1007/s11136-014-0652-5.

[177] Zachariou A, Filiponi M, Kirana PS. Translation and validation of the Greek version of the female sexual function index questionnaire. International Journal of Impotence Research 2017;29:171–4. https://doi.org/10.1038/ijir.2017.18.

[178] Zohre M, Minoo P, Ali M. Pelvic Organ Prolapse/Urinary Incontinence Sexual Questionnaire (PISQ-12): psychometric validation of the Iranian version. International Urogynecology Journal and Pelvic Floor Dysfunction 2014;26:433–9. https://doi.org/10.1007/s00192-014-2520-9.

[179] Chedraui P, Pérez-López FR, Sánchez H, Aguirre W, Martínez N, Miranda O, et al. Assessment of sexual function of mid-aged Ecuadorian women with the 6-item Female Sexual Function Index. Maturitas 2012;71:407–12. https://doi.org/10.1016/j.maturitas.2012.01.013.

[180] Lee Y, Lim MC, Joo J, Park K, Lee S, Seo S, et al. Development and Validation of the Korean Version of the Female Sexual Function Index-6 (FSFI-6K). Yonsei Medical Journal 2014;55:1442–6. https://doi.org/10.3349/YMJ.2014.55.5.1442.

[181] Mitchell KR, Ploubidis GB, Datta J, Wellings K. The Natsal-SF: A validated measure of sexual function for use in community surveys. European Journal of Epidemiology 2012;27:409–18. https://doi.org/10.1007/s10654-012-9697-3.

[182] Pérez-López FR, Fernández-Alonso AM, Trabalón-Pastor M, Vara C, Chedraui P. Assessment of sexual function and related factors in mid-aged sexually active Spanish women with the six-item Female Sex Function Index. Menopause 2012;19:1224–30. https://doi.org/10.1097/gme.0b013e3182546242.

[183] Bartula I, Sherman KA. Development and validation of the Female Sexual Function Index adaptation for breast cancer patients (FSFI-BC). Breast Cancer Research and Treatment 2015;152:477–88. https://doi.org/10.1007/s10549-015-3499-8.

[184] Burri A, Cherkas L, Spector T. Replication of psychometric properties of the FSFI and validation of a modified version (FSFI-LL) assessing lifelong sexual function in an unselected sample of females. Journal of Sexual Medicine 2010;7:3929–39. https://doi.org/10.1111/j.1743-6109.2010.01970.x.

[185] Carvalho J, Vieira AL, Nobre P. Latent Structures of Female Sexual Functioning. Archives of Sexual Behavior 2012;41:907–17. https://doi.org/10.1007/s10508-011-9865-7.

[186] Bland MJ, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet 1986;327:307–10. https://doi.org/10.1016/S0140-6736(86)90837-8.

[187] Bland JM, Altman DG. Agreement between methods of measurement with multiple observations per individual. Journal of Biopharmaceutical Statistics 2007;17:571–82. https://doi.org/10.1080/10543400701329422.

[188] Carvalheira AA, Brotto LA, Leal I. Women's Motivations for Sex: Exploring the Diagnostic and Statistical Manual, Fourth Edition, Text Revision Criteria for Hypoactive Sexual Desire and Female Sexual Arousal Disorders. The Journal of Sexual Medicine 2010;7:1454–63. https://doi.org/10.1111/j.1743-6109.2009.01693.x.

[189] Balon R, Segraves RT, Clayton A. Issues for DSM-V: Sexual Dysfunction, Disorder, or Variation Along Normal Distribution: Toward Rethinking DSM Criteria of Sexual Dysfunctions. American Journal of Psychiatry 2007;164:198–200. https://doi.org/10.1176/ajp.2007.164.2.198.

[190] Gierhart BS. When does a "less than perfect" sex life become female sexual dysfunction? Obstetrics & Gynecology 2006;107:750–1.

[191] Sungur MZ, Gündüz A. A Comparison of DSM-IV-TR and DSM-5 Definitions for Sexual Dysfunctions: Critiques and Challenges. The Journal of Sexual Medicine 2014;11:364–73. https://doi.org/10.1111/jsm.12379.

[192] Vet HC de, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health and Quality of Life Outcomes 2006;4:54. https://doi.org/10.1186/1477-7525-4-54.

[193] Puppo V, Puppo G. RE: Bartula I, Sherman KA: Development and validation of the Female Sexual Function Index adaptation for breast cancer patients (FSFI-BC). Breast Cancer Research and Treatment 2015;153:705–6. https://doi.org/10.1007/s10549-015-3574-1.

[194] Browne K, Roseman D, Shaller D, Edgman-Levitan S. Analysis & commentary: Measuring patient experience as a strategy for improving primary care. Health Affairs 2010;29:921–5. https://doi.org/10.1377/hlthaff.2010.0238.

[195] Brédart A, Bottomley A. Treatment satisfaction as an outcome measure in cancer clinical treatment trials. Expert Review of Pharmacoeconomics & Outcomes Research 2002;2:597-606. https://doi.org/10.1586/14737167.2.6.597.

[196] Bjordal K, Graeff A de, Fayers PM, Hammerlid E, Pottelsberghe C van, Curran D, et al. A 12 country field study of the EORTC QLQ-C30 (version 3.0) and the head and neck cancer specific module (EORTC QLQ-H&N35) in head and neck patients. European Journal of Cancer 2000;36:1796–807. https://doi.org/10.1016/S0959-8049(00)00186-6.

[197] Neijenhuijs K, Verdonck-de Leeuw I, Cuijpers P, Hout A van der, Melissant H, Wit M de, et al. Validity and reliability of patient reported outcomes measuring quality of life in cancer patients 2017.

[198] Hjörleifsdóttir E, Hallberg IR, Gunnarsdóttir ED. Satisfaction with care in oncology outpatient clinics: psychometric characteristics of the Icelandic EORTC IN-PATSAT32 version. Journal of Clinical Nursing 2010;19:1784–94. https://doi.org/10.1111/j.1365-2702.2009.03095.x.

[199] Pishkuhi MA, Salmaniyan S, Nedjat S, Zendedel K, Lari MA. Psychometric properties of the Persian version of satisfaction with care EORTC-in-patsat32 questionnaire among Iranian cancer patients. Asian Pacific Journal of Cancer Prevention 2014;15:10121–8. https://doi.org/10.7314/APJCP.2014.15.23.10121.

[200] Arraras JI, Vera R, Martínez M, Hernández B, Laínez N, Rico M, et al. The EORTC cancer in-patient satisfaction with care questionnaire: EORTC IN-PATSAT32. Clinical and Translational Oncology 2009;11:237–42. https://doi.org/10.1007/s12094-009-0346-6.

[201] Zhang J, Xie S, Liu J, Sun W, Guo H, Hu Y, et al. Validation of EORTC IN-PATSAT32 for Chinese patients with gastrointestinal cancer. Patient Preference Adherence 2014;8:1285–92. https://doi.org/10.2147/ppa.s67111.

[202] Zhang L, Dai Z, Cheng S, Xie S, Woo SML, Luo Z, et al. Validation of EORTC IN-PATSAT32 for Chinese cancer patients. Supportive Care Cancer 2015;23:2721–30. https://doi.org/10.1007/s00520-015-2636-6.

[203] Obtel M, Serhier Z, Bendahhou K. Validation of EORTC IN-PATSAT 32 in Morocco: Methods and Processes. Asian Pacific Journal 2017;18:1403–9. https://doi.org/10.22034/APJCP.2017.18.5.1403.

[204] Cortina JM. What is coefficient alpha? An examination of theory and applications. Journal of Applied Psychology 1993;78:98–104. https://doi.org/10.1037/0021-9010.78.1.98.

[205] Asadi-lari M, Ahmadi Pishkuhi M, Almasi-Hashiani A, Safiri S, Sepidarkish M. Validation study of the EORTC information questionnaire (EORTC QLQ-INFO25) in Iranian cancer patients. Supportive Care Cancer 2015;23:1875–82. https://doi.org/10.1007/s00520-014-2510-y.

[206] Arraras JI, Greimel E, Sezer O, Chie WC, Bergenmar M, Costantini A, et al. An international validation study of the EORTC QLQ-INFO25 questionnaire: an instrument to assess the information given to cancer patients. European Journal of Cancer 2010;46:2726–38. https://doi.org/10.1016/j.ejca.2010.06.118.

[207] Groenvold M, Petersen MA, Aaronson NK, Arraras JI, Blazeby JM, Bottomley A, et al. The development of the EORTC QLQ-C15-PAL: A shortened questionnaire for cancer patients in palliative care. European Journal of Cancer 2006;42:55–64. https://doi.org/10.1016/j.ejca.2005.06.022.

[208] Aboshaiqah A, Al-Saedi TSB, Abu-Al-Ruyhaylah MMM, Aloufi AA, Alharbi MO, Alharbi SSR, et al. Quality of life and satisfaction with care among palliative cancer patients in Saudi Arabia. Palliative and Supportive Care 2016;14:621–7. https://doi.org/10.1017/S1478951516000432.

[209] Neijenhuijs KI. Measurement Error in Psychological Science 2019.

[210] Wang Y-P, Gorenstein C. Psychometric properties of the Beck Depression Inventory-II: a comprehensive review. Brazilian Journal of Psychiatry 2013;35:416–31.

[211] Posternak M, Miller I. Untreated short-term course of major depression: a meta-analysis of outcomes from studies using wait-list control groups. Journal of Affective Disorders 2001;66:139–46.

[212] Button KS, Kounali D, Thomas L, Wiles NJ, Peters TJ, Welton NJ, et al. Minimal clinically important difference on the Beck Depression Inventory–II according to the patient's perspective. Psychological Medicine 2015;45:3269–79. https://doi.org/10.1017/S0033291715001270.

[213] Muchinsky PM. The Correction for Attenuation. Educational and Psychological Measurement 1996;56:63–75. https://doi.org/10.1177/0013164496056001004.

[214] Rouder J, Kumar A, Haaf JM. Why Most Studies of Individual Differences With Inhibition Tasks Are Bound To Fail. Pre-Print at PsyArXiv 2019. https://doi.org/10.31234/OSF.IO/3CJR5.

[215] Ozonoff D. Statistics in peer review 2006. https://doi.org/10.1038/nature04989.

[216] Elbert NJ, Os-Medendorp H van, Renselaar W van, Ekeland AG, Hakkaart-van Roijen L, Raat H, et al. Effectiveness and cost-effectiveness of ehealth interventions in somatic diseases: a systematic review of systematic reviews and meta-analyses. Journal of Medical Internet Research 2014;16:e110. https://doi.org/10.2196/jmir.2790.

[217] Slev VN, Mistiaen P, Pasman HRW, Leeuw IMV-d, Uden-Kraan CF van, Francke AL. Effects of eHealth for patients and informal caregivers confronted with cancer: A meta-review. International Journal of Medical Informatics 2016;87:54–67. https://doi.org/10.1016/J.IJMEDINF.2015.12.013.

[218] Eurostat. Households level of internet access n.d.

[219] Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Quality of Life Research 2011;20:1727–36. https://doi.org/10.1007/s11136-011-9903-x.

[220] Ishikawa H, Takeuchi T, Yano E. Measuring functional, communicative, and critical health literacy among diabetic patients. Diabetes Care 2008;31:874–9. https://doi.org/10.2337/dc07-1932.

[221] Brooks J. SUS: A "quick and dirty" usability scale. In: Jordan P, Thomas B, Weerdmeester B, McClelland I, editors. Usability evaluation in industry, London, UK: Taylor & Francis; 1996, pp. 189–94.

[222] R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2019.

[223] Rosseel Y. lavaan: An R package for structural equation modeling. Journal of Statistical Software 2012;48:1–36.

[224] Jorgensen TD, Pornprasertmanit S, Schoemann AM, Rosseel Y. `semTools`: Useful tools for structural equation modeling. 2019.

[225] McDonald RP. Test theory: A unified treatment mcdonald - Google Scholar. Mahwah, NJ: Erlbaum. 1999.

[226] Gamer M, Lemon J, <puspendra.pusp22@gmail.com> IFPS. Irr: Various coefficients of interrater reliability and agreement. 2019.

[227] Fletcher TD. Psychometric: Applied psychometric theory. 2010.

[228] Harrell Jr FE, Charles Dupont, others. Hmisc: Harrell miscellaneous. 2019.

[229] Rizopoulos D. Ltm: An r package for latent variable modelling and item response theory analyses. Journal of Statistical Software 2006;17:1–25.

[230] Irwin MR, Olmstead RE, Ganz PA, Haque R. Sleep disturbance, inflammation and depression risk in cancer survivors. Brain, Behavior, and Immunity 2013;30:S58–67. https://doi.org/10.1016/J.BBI.2012.05.002.

[231] Cuijpers P, Beekman A, Smit F, Deeg D. Predicting the onset of major depressive disorder and dysthymia in older adults with subthreshold depression: a community based study. International Journal of Geriatric Psychiatry 2006;21:811–8. https://doi.org/10.1002/gps.1565.

[232] Pullens MJJ, De Vries J, Roukema JA. Subjective cognitive dysfunction in breast cancer patients: a systematic review. Psycho-Oncology 2010;19:1127–38. https://doi.org/10.1002/pon.1673.

[233] Alfano C, Rowland J. Recovery issues in cancer survivorship: a new challenge for supportive care. The Cancer Journal 2006;12:432–43.

[234] Campello RJGB, Moulavi D, Sander J. Density-Based Clustering Based on Hierarchical Density Estimates. In:. Pacific-asia conference on knowledge discovery and data mining, Berlin, Heidelberg: Springer; 2013, pp. 160–72. https://doi.org/10.1007/978-3-642-37456-2_14.

[235] Duineveld LAM, Wieldraaijer T, Asselt KM van, Nugteren IC, Donkervoort SC, Ven AWH van de, et al. Improving care after colon cancer treatment in The Netherlands, personalised care to enhance quality of life (I CARE study): study protocol for a randomised controlled trial. Trials 2015;16:284. https://doi.org/10.1186/s13063-015-0798-7.

[236] Python Software Foundation. Python Language Reference 2018.

[237] Mcinnes L, Healy J, Astels S. hdbscan: Hierarchical density based clustering. The Open Journal 2017;2. https://doi.org/10.21105/joss.00205.

[238] Ester M, Kriegel H, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In:. Proceedings of the second international conference on knowledge discovery and data mining (kdd-96), AAAI Pres; 1996. https://doi.org/10.1.1.121.9220.

[239] Pedersen TL. Tidygraph: A tidy api for graph manipulation. 2019.

[240] Pedersen TL. Ggraph: An implementation of grammar of graphics for graphs and networks. 2019.

[241] Vandenberg RJ, Lance CE. A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. Organizational Research Methods 2000;3:4–70. https://doi.org/10.1177/109442810031002.

[242] Schmitt N, Kuljanin G. Measurement invariance: Review of practice and implications. Human Resource Management Review 2008;18:210–22. https://doi.org/10.1016/J.HRMR.2008.03.003.

[243] Sousa VD, Rojjanasrirat W. Translation, adaptation and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline. Journal of Evaluation in Clinical Practice 2011;17:268–74. https://doi.org/10.1111/j.1365-2753.2010.01434.x.

[244] Davis-Stober CP, Dana J, Rouder JN. Estimation accuracy in the psychological sciences. PLOS ONE 2018;13:e0207239. https://doi.org/10.1371/journal.pone.0207239.

[245] Vickers AJ. Validation of Patient-Reported Outcomes: A Low Bar. Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology 2019:JCO1901126. https://doi.org/10.1200/JCO.19.01126.

[246] Jöreskog KG. A general approach to confirmatory maximum likelihood factor analysis. Psychometrika 1969;34:183-202. https://doi.org/10.1007/BF02289343.

[247] Bartko JJ. The Intraclass Correlation Coefficient as a Measure of Reliability. Psychological Reports 1966;19:3–11. https://doi.org/10.2466/pr0.1966.19.1.3.

[248] Heise D, Bohrnstedt G. Validity, invalidity, and reliability. Sociological Methodology 1970;2:104–29.

[249] Cole DA. Utility of Confirmatory Factor Analysis in Test Validation Research. Journal of Consulting and Clinical Psychology 1987;55:584–94. https://doi.org/10.1037/0022-006X.55.4.584.

[250] Ferketich S. Internal consistency estimates of reliability. Research in Nursing & Health 1990;13:437–40. https://doi.org/10.1002/nur.4770130612.

[251] Bryant F, Yarnold P. Principal-components analysis and exploratory and confirmatory factor analysis. In: Grimm LG, Yarnold PR, editors. Reading and understanding multivariate statistics, Washington, DC: American Psychological Associatio; 1995, pp. 99–136.

[252] Donner A. A Review of Inference Procedures for the Intraclass Correlation Coefficient in the One-Way Random Effects Model. International Statistical Review / Revue Internationale de Statistique 1986;54:67. https://doi.org/10.2307/1403259.

[253] Müller R, Büttner P. A critical discussion of intraclass correlation coefficients. Statistics in Medicine 1994;13:2465–76. https://doi.org/10.1002/sim.4780132310.

[254] Hout A van der, Van Uden-Kraan CF, Holtmaat K, Verdonck-de Leeuw IM. Role of eHealth application Oncokompas in supporting self-management of symptoms and health-related quality of life in cancer survivors: a randomised controlled trial. Lancet Oncology 2019.

[255] Open Science Collaboration OS. Estimating the reproducibility of psychological science. Science 2015;349:aac4716–6. https://doi.org/10.1126/science.aac4716.

[256] Klein RA, Vianello M, Hasselman F, Adams BG, Adams RB, Alper S, et al. Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. Advances in Methods and Practices in Psychological Science 2018;1:443–90. https://doi.org/10.1177/2515245918810225.

[257] Begley CG, Ioannidis JPA. Reproducibility in Science. Circulation Research 2015;116:116–26. https://doi.org/10.1161/CIRCRESAHA.114.303819.

[258] Maxwell SE, Lau MY, Howard GS. Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? American Psychologist 2015;70:487–98. https://doi.org/10.1037/a0039400.

[259] Colquhoun D. The reproducibility of research and the misinterpretation of p-values. Royal Society Open Science 2017;4:171085. https://doi.org/10.1098/rsos.171085.

[260] Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, et al. Redefine statistical significance. Nature Human Behaviour 2018;2:6–10. https://doi.org/10.1038/s41562-017-0189-z.

[261] King G, Crosas M. The Dataverse Project n.d.

[262] Crosas M. The Dataverse Network: An Open-Source Application for Sharing, Discovering and Preserving Data. D-Lib Magazine 2011;17. https://doi.org/10.1045/january2011-crosas.

[263] Kauppinen T. Linked Scientist n.d.

[264] OSF. Open Science Framework n.d.

[265] Foster, MSLS ED, Deardorff, MLIS A. Open Science Framework (OSF). Journal of the Medical Library Association 2017;105:203. https://doi.org/10.5195/JMLA.2017.88.

[266] Open Science Collaboration. Estimating the reproducibility of psychological science. Science 2015;349:aac4716–6. https://doi.org/10.1126/science.aac4716.

[267] Chandler J, Shapiro D. Conducting Clinical Research Using Crowdsourced Convenience Samples. Annual Review of Clinical Psychology 2016;12:53–81. https://doi.org/10.1146/annurev-clinpsy-021815-093623.

[268] Amazon. Amazon Mechanical Turk n.d.

[269] Berinsky AJ, Huber GA, Lenz GS. Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. Political Analysis 2012;20:351–68. https://doi.org/10.1093/pan/mpr057.

[270] Clickworker. Data Management Services: AI training data, text creation, web researches n.d.

[271] MicroWorkers. Microworkers - work & earn or offer a micro job n.d.

[272] Science Europe. 'Plan S' and 'cOAlition S' – Accelerating the transition to full and immediate Open Access to scientific publications 2019.

[273] Stocker M. From Data to Machine Readable Information Aggregated in Research Objects. D-Lib Magazine 2017;23. https://doi.org/10.1045/january2017-stocker.