

LONDON
SCHOOL of
HYGIENE
& TROPICAL
MEDICINE



Protopapa, E.; van der Meulen, J.; Moore, C.M.; Smith, S.C. (2017) [Accepted Manuscript] Patient Reported Outcome (PRO) questionnaires for men who have radical surgery for prostate cancer: a conceptual review of existing instruments. *BJU international*. ISSN 1464-4096 DOI: <https://doi.org/10.1111/bju.13896>

Downloaded from: <http://researchonline.lshtm.ac.uk/3962321/>

DOI: [10.1111/bju.13896](https://doi.org/10.1111/bju.13896)

Usage Guidelines

Please refer to usage guidelines at <http://researchonline.lshtm.ac.uk/policies.html> or alternatively contact researchonline@lshtm.ac.uk.

Available under license: <http://creativecommons.org/licenses/by-nc-nd/2.5/>

MS. EVANGELIA PROTOPAPA (Orcid ID : 0000-0002-1597-1462)

Article type : Review

Article Category: Review

Patient Reported Outcome (PRO) questionnaires for men who have radical surgery for prostate cancer: a conceptual review of existing instruments

Evangelia Protopapa MSc,¹ Jan van der Meulen PhD FFPH,¹ Caroline M. Moore MD FRCS (Urol),² Sarah C. Smith PhD,¹

¹Department of Health Services Research & Policy, London School of Hygiene & Tropical Medicine

²Division of Surgery and Interventional Science, University College London

Corresponding author:

Sarah Smith

Department of Health Services Research & Policy

London School of Hygiene & Tropical Medicine

15-17 Tavistock Place, London, UK

Telephone: +44 207 927 2038

Email: sarah.smith@lshtm.ac.uk

Co-authors' contact details:

evangelia.protopapa@lshtm.ac.uk

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/bju.13896

jan.vandermeulen@lshtm.ac.uk

caroline.moore@ucl.ac.uk

Funding sources: This work is funded by Prostate Cancer UK. The funder had no role in any of the following: design and conduct of the study, data collection and management, data analysis and interpretation, or preparation, approval and review of the manuscript.

Financial Disclosure: The authors have no financial relationships relevant to this article to disclose.

Potential Conflicts of Interest: The authors have no conflicts of interest relevant to this article to disclose.

Key words: patient-reported outcomes, prostate cancer, health-related quality of life, psychometric

Abstract

Objectives: To critically review conceptual frameworks for available patient reported outcome (PRO) questionnaires in men having radical prostatectomy; psychometrically evaluate each questionnaire; identify whether each is appropriate for use at the level of the individual patient.

Materials and Methods: We searched PubMed, the Reports and Publications database of the University of Oxford Patient Reported Outcomes Measurement Group and the website

of the International Consortium for Health Outcomes Measurement (ICHOM) for psychometric reviews of prostate cancer specific PRO questionnaires. From these we identified relevant questionnaires and critically appraised the conceptual content, guided by the Wilson and Cleary framework and psychometric properties, using well established criteria.

Results: Searches found four reviews and one recommendation paper. We identified seven prostate cancer specific PROs (EPIC-26, EPIC-50, UCLA-PCI, FACT-P, QLQ-PR25, and PC-QoL and STAR). Six out of seven measures purported to measure health related quality of life, but items focused strongly on urinary and sexual symptoms/functioning. The remaining questionnaire (STAR) claimed to assess functional recovery after radical prostatectomy. The psychometric evidence for these questionnaires was incomplete and variable in quality; none had evidence that they were appropriate for use with individual patients.

Conclusion: Several questionnaires provide the basis of measures of urinary and/or sexual symptoms/functioning. Further work should explore other aspects of health related quality of life that are important for men having radical prostatectomy. Further psychometric work is also needed to determine whether they can be used at the individual level.

INTRODUCTION

Patient reported outcomes (PROs) are increasingly being used. Patients' reports of symptoms, function and health-related quality of life (HRQL) or overall quality of life (QoL) can provide important information about the impact that health problems and related treatments have on patients' lives [1, 2]. These terms are often used interchangeably and

for the remainder of this paper we will use the term HRQL. Typically, PROs are used to study outcomes at group level. However, in some prostate cancer centres, a formal tool for the collection of PROs has been implemented in routine clinical practice to monitor outcomes in individual patients after radical surgery [3].

In the UK, around 10% of men newly diagnosed with prostate cancer have a radical prostatectomy [4]. In addition to cancer cure, surgeons consider the absence of side effects, such as urinary incontinence and erectile dysfunction, as important indicators of the success of the surgery. Men themselves are interested in a wider range of outcomes [5], including anxiety and distress, social interactions and intimate relations [6], feelings of masculinity, and self-esteem [7]. It is therefore important that both developers and users of PRO questionnaires for men diagnosed with prostate cancer are clear about what the questionnaire is supposed to measure, whether the questionnaire actually measures these concepts, and how well it does so.

Several reviews have evaluated the scientific or psychometric properties of existing PRO questionnaires that are being used for patients diagnosed with prostate cancer [8-10]. In general, these reviews have used well-established guidelines and criteria based on classical-test theory to evaluate the robustness of existing instruments [11-13]. The performance of these instruments is often expressed in terms of reliability (the extent to which an instrument is free from error), validity (the extent to which it measures what it aims to measure), and responsiveness (the ability to detect clinically important changes over time). However, none of the existing reviews has included a critical review of the conceptual content of existing instruments nor evaluated the extent to which these instruments are fit to monitor outcomes in individual patients.

To evaluate instruments' fitness to measure outcomes at the individual level, modern psychometric approaches are now often used, such as Rasch measurement theory (RMT) and item response theory (IRT) [14-16]. Scores derived from these approaches have a number of advantages over methods based on classical-test theory: they are truly "interval scores", meaning that there is an equal distance between each of the values on the scale (as opposed to ordinal scores where values are in rank order but the distance between any two

values may not be equal), “invariant” (independent of both the distribution of items and the distribution of the sample), and potentially appropriate for use in individual patients.

Inconsistency about how HRQL is defined adds to the complexity of evaluating PRO questionnaires that claim to measure these constructs. Although there is no universal definition of HRQL, it is generally agreed that it is multi-dimensional and subjective [17-19]. That is, HRQL focuses on the perceived impact (from the patient’s perspective) of physical, mental and social domains of health. HRQL is therefore not usually concerned with how much of a symptom a patient has, but rather to what extent the patient is “bothered by” or “concerned about” that symptom. In addition, Wilson and Cleary [20] have provided a conceptual framework that places HRQL in the wider context of health outcomes and suggests how HRQL is related to other health outcomes widely used in clinical and health services research. They distinguish five levels of outcome: “biological and physiological variables”, “symptom status”, “functional status”, “general health perceptions”, and “overall QoL”. These five outcomes are considered to be separate constructs that are causally related (e.g. “symptom status” will affect “functional status” which in turn is likely to affect HRQL). It is important to note that the “overall QoL” construct in the Wilson and Cleary framework should reflect individuals’ subjective perception of how happy or satisfied they are with their life as a whole.

To address the shortcomings of previous reviews, and to evaluate whether the HRQL outcomes that are important to men are being assessed, we carried out a review of the existing psychometric reviews of prostate cancer-specific instruments developed to collect PROs. Our aims were to 1) critically review the conceptual content of available PRO questionnaires used in men having radical prostatectomy, 2) psychometrically evaluate each instrument and 3) determine the extent to which each instrument is appropriate for use in individual patients. We used this review-of-reviews to identify the original development articles and we applied the Wilson and Cleary framework as the basis for a critical appraisal of the instruments’ conceptual content [20]. We also summarised the available psychometric evidence for each instrument and assessed the extent to which each instrument is appropriate for use in individual patients.

METHODS

Search strategy

We searched the PubMed database from its beginning to June 2015 to identify psychometric reviews of prostate cancer-specific PRO questionnaires. We searched PubMed using the following search strategy: (“quality of life” OR “QoL” OR “HRQL” OR “symptom” OR “function” OR “disability” OR “patient reported outcome”) AND (“prostate”) AND (“instrument” OR “measure ” OR “questionnaire”), limited by “review” and “systematic review”.

To identify additional reviews we searched the Reports and Publications database of the University of Oxford Patient Reported Outcomes Measurement Group [21]. We also searched the website of the International Consortium for Health Outcomes Measurement (ICHOM) for reports of recommendations on prostate cancer-specific PROs [21].

Instrument selection

From the identified review papers, we compiled a list of instruments recommended by the reviewers following psychometric assessment, and located the original development articles. We excluded instruments developed in languages other than English, generic cancer-related instruments, instruments that were designed to capture utilities (i.e. quantitative measures of individuals’ preferences for specific health states), instruments not developed for use in patients with prostate cancer, and instruments specifically designed to be used in men who had prostate cancer treatments other than radical prostatectomy. Single question assessments were also excluded.

Data extraction and appraisal

We identified the original development article(s) for each instrument, and extracted data for each instrument, using a standard data extraction form derived from Smith et al 2005 [22]. The criteria used for this appraisal are described in Table 1 and based on well-established classical psychometric criteria [11, 12, 23]. We chose these guidelines because

they were specifically focused on the criteria for the psychometric properties that an instrument must have, rather than the quality of the paper reporting the instrument development, which the COSMIN guidelines address [24]. To assess content validity we also undertook a conceptual review of each of the identified instruments using the five levels of outcome included in the Wilson and Cleary framework [20]. We used this framework as a guide to compare the items of each instrument with the construct that each instrument claimed to measure. For example, if an instrument claimed to measure HRQL, we expected items to reflect the subjectively perceived impact of physical, mental and social domains of health rather than the objective level of physical, mental or social health. That is, an item measuring HRQL would ask about the extent to which a man is concerned or bothered by his symptoms rather than the extent to which he has symptoms. Two authors (SCS, EP) reviewed each instrument using the standard extraction form. These two authors completed the review independently and then discussed any discrepancies until reaching consensus.

RESULTS

The literature search generated 450 review papers. 447 articles were identified via PubMed, while three articles were identified from the remaining sources. After removing duplicates, a total of 448 abstracts were screened and 33 full-text articles reviewed. Four papers presenting the results of psychometric reviews [8-10, 25] were identified and one paper describing recommendations by ICHOM [26] (Figure 1).

Based on the recommendations made by these five papers, we identified six prostate cancer-specific PRO questionnaires for further review (EPIC-26, EPIC-50, UCLA-PCI, FACT-P, QLQ-PR25, and PC-QoL) and located the original development article for each [27-35]. FACT-P [33] and EORTC QLQ-C30 [36] are generic cancer-related QoL instruments with an additional prostate cancer-specific module. The prostate cancer specific module of EORTC QLQ-C30 is known as QLQ-PR25 [34]. We evaluated only the prostate cancer specific

subscales for these two PRO questionnaires. An instrument, developed by Memorial Sloan Kettering Cancer Center (New York, US), known as STAR [3], to monitor PROs following radical prostatectomy in routine clinical practice is also included in this review because it is widely used, and was developed specifically for repeat use in men having radical prostatectomy. Therefore, seven instruments are described in Table 2.

Conceptual Review

Our conceptual review addressed the content validity of each instrument in terms of the extent to which the items in each instrument reflected the construct that the developers claimed the instrument measured. The developers of six of the seven identified instruments themselves used the terms “HRQL” or “overall QL” to describe what their instruments measured (EPIC-26, EPIC-50, UCLA-PCI, EORTC QLQ-PR25, and PC-QoL).¹ The remaining instrument, (STAR) aimed to measure “functional recovery” after radical prostatectomy.

The results of our conceptual review are shown in Figures 2 to 8. For each instrument, we allocated each item (question) to one of Wilson & Cleary’s levels of outcome (i.e. “biological and physiological variables”, “symptom status”, “functional status”, “general health perceptions”, and “overall QoL”). This allowed us to critique the conceptual content of each instrument by determining which level of outcome each item represented. None of the seven instruments include items related to the Wilson and Cleary outcome level labelled as “biological and physiological variables”. All instruments include items related to “symptom status” or “functional status”, particularly sexual and urinary problems and all instruments except EPIC-26 include items about bowel problems. In addition, EPIC-50 and EORTC-QLQ also include hormonal symptoms and FACT-P includes other symptoms and functional problems, such as weight loss, appetite and pain.

STAR is the only instrument that includes an item related to “general health perceptions”, asking about the overall feeling related to the current state of health.

¹ Please note that these are the terms are used by the developers. They often did not explicitly define what these terms included and do not necessarily reflect the definitions used in the conceptual framework proposed by Wilson and Cleary.¹⁴

All instruments included some items that can be labelled as “QoL”. Across all instruments these are mostly focused on the subjective impact of physical health, with fewer items representing the perceived impact of mental health. EPIC-26, EPIC-50 and PC-QOL include the most items that can be labelled as “QoL”. In general, most of these items reflect the subjective impact of physical aspects of health (e.g. the extent to which dripping or leaking urine is perceived as a problem or the extent to which changes in body weight are perceived as a problem). PC-QOL also includes items that reflect the impact of mental aspects of health (e.g. concern/anxiety about treatment, recurrence, and quality of care). However, items reflecting the subjective impact of social aspects of health were less common. Only PC-QOL includes an item about worry arising from being unable to please a partner sexually, an aspect of social health.

In the items that could be labelled as Wilson & Cleary’s term “overall QOL” (though this was often described as HRQL by developers of the questionnaires) in the remaining instruments, there was a similar predominance of items asking about the subjective impact of physical aspects of health. The HRQL items in QLQ-PR25 include three items about the subjective impact of physical health and one about the impact of mental health (feeling less masculine). Similarly, the HRQL items in UCLA-PCI include four items about the subjective impact of aspects of physical health (e.g. the extent to which weight gain, weight loss and incontinence aids have been a problem) and one item about the impact of mental health (feeling less masculine as a result of treatment). STAR and FACT-P PCS have the narrowest HRQL focus. STAR includes two HRQL items, both assessing the subjective impact of aspects of physical health (the extent to which urinary and bowel function are a problem). FACT-P PCS includes three HRQL items, of which two assess the subjective impact of physical aspects of health (satisfaction with levels of pain and comfort level) and one assesses the impact of mental health (ability to feel like a man).

Psychometric appraisal

The detailed psychometric review of the seven identified instruments, based on their development papers is described below, and the results of their psychometric appraisal is

shown in Table 3. Overall, the psychometric evidence supporting the instruments was patchy and variable in quality.

Classical test theory

Evidence supporting instrument acceptability was weak for all instruments except EPIC-50, with high floor and/or ceiling effects across most instruments. All instruments had evidence supporting their reliability, although this evidence was weak for STAR and FACT-P PCS.

These scales (together with QLQ-PR25) only assessed one type of reliability (internal consistency).

For all instruments there was evidence of at least one other form of validity. This evidence was weakest for PCS (FACT-P) which only had weak evidence for known groups differences.

Validity evidence was moderately strong for the other instruments, across a range of different types of validity.

Lastly, evidence regarding the responsiveness of all instruments was very limited which makes it impossible to assess the instruments' ability to detect clinically important differences in HRQL in relation to treatment over time.

Modern test theory

No instrument was developed or subsequently analysed using item response theory or Rasch measurement theory.

CONCLUSION

We found that the developers of most of the seven identified prostate-cancer specific PROs claim that these instruments measure "HRQL" or "overall QoL", but their items strongly focus on urinary, sexual and bowel symptoms and function. All questionnaires include some items with a more subjective element to determine the extent to which men are concerned or bothered by a particular symptom. However, no instrument includes the full range of

Accepted Article

items necessary to represent HRQL in terms of the subjective impact of physical, mental and social aspects of prostate cancer. These gaps do not only compromise the content validity of the six instruments that claim to measure HRQL but they also affect the interpretation of the scores and their suitability for use in research and service evaluation. This means that in clinical practice the true impact of prostate cancer treatment is not reflected in the current outcomes. The currently available instruments do not measure the range of outcomes that are important to men.

Generally, evidence of reliability and validity is incomplete for all instruments. EPIC-50, EPIC-26 and UCLA-PCI and PC-QOL have most evidence for robust psychometric properties. Of these, EPIC-26 has the advantage that it is well-used and has comparable psychometric properties to EPIC-50, but is considerably shorter. However, as none of the instruments has evidence of responsiveness it is impossible to draw any conclusion about whether they are sensitive to clinically meaningful change. None of the instruments has been evaluated using modern psychometric methods and consequently we have no evidence about how well they work at the individual patient level.

Although only PC-QOL had content validity as a measure of HRQL (i.e. it included items reflecting the perceived impact of physical, mental and social aspects of health), closer inspection of the items across all of the instruments suggests that most reflect the recognised side effects of surgery. However, the relatively narrow focus and dearth of items reflecting patients' subjective feelings about the impact of physical, mental and social aspects of health means that they may not reflect all the outcomes that are important to patients. There is a need for greater qualitative understanding of HRQL for men with prostate cancer and to develop questionnaire items that reflect this. Additionally, further psychometric development work is needed, using modern psychometric methods (such as IRT or RMT) to determine the extent to which it is appropriate to use these questionnaires at the individual level.

Our findings are limited by the fact that we used the instruments that were recommended by other reviews. The five reviews overlapped in their recommendations, especially for EPIC-50, EPIC-26 and UCLA-PCI. As these recommendations are often the basis of how

researchers and practitioners choose instruments, it is an appropriate shortlist for further critique. The authors of each review used slightly different psychometric criteria to reach their conclusions, but there was also much similarity in the criteria that were used.

In addition, the review of questionnaire items and their conceptual content involves a degree of judgement. We have made our definitions explicit and based them on the widely available and often cited literature in this area. The categorisation of each item was undertaken by two of the authors and later also reviewed by the remaining authors.

The results of our review, and especially the evaluation of conceptual content, suggest that the available PROs offer a limited evaluation of the outcomes after radical prostatectomy that are relevant to men with prostate cancer. There are gaps in content and also inadequate evidence of reliability, validity and responsiveness of the existing instruments.

In conclusion, several instruments provide the basis of measures of urinary symptoms and/or sexual function that could potentially be used at the group level. Although the focus on symptoms and functional outcomes is of interest to both clinicians and patients, there are other aspects of HRQL that need to be explored as important outcomes for men receiving surgery for prostate cancer. As yet, there is no formal evidence to support the appropriateness of the questionnaires for use at the individual level.

Acknowledgments

This work is part of the True NTH UK Post Surgery project. True NTH is a global initiative by the Movember Foundation which aims to identify and demonstrate the best and most cost-effective models for improving prostate cancer survivorship care and support.

Conflicts of Interest

None

References

1. Chard J, Kuczawski M, Black N, van der Meulen J. Outcomes of elective surgery undertaken in independent sector treatment centres and NHS providers in England: audit of patient outcomes in surgery. *BMJ* 2011; 343.
2. Browne J, Jamieson L, Lewsey J, van der Meulen J, Black N, Cairns J, et al. Patient reported outcome measures (PROMs) in elective surgery. *Report to the Department of Health* 2007; 12.
3. Vickers AJ, Savage CJ, Shouery M, Eastham JA, Scardino PT, Basch EM. Validation study of a web-based assessment of functional recovery after radical prostatectomy. *Health and Quality of Life Outcomes* 2010; 8: 82.
4. National Cancer Intelligence Network. Treatment routes in prostate cancer 2012 [May 2015]. Available from: <http://www.ncin.org.uk/view.aspx?rid=1260>.
5. Glaser AW, Fraser LK, Corner J, Feltbower R, Morris EJ, Hartwell G, et al. Patient-reported outcomes of cancer survivors in England 1-5 years after diagnosis: a cross-sectional survey. *BMJ Open* 2013; 3.
6. O'Shaughnessy P, Laws TA. Australian men's long term experiences following prostatectomy: a qualitative descriptive study. *Contemp Nurse* 2009; 34: 98-109.
7. Hedestig O, Sandman PO, Tomic R, Widmark A. Living after radical prostatectomy for localized prostate cancer: a qualitative analysis of patient narratives. *Acta Oncol* 2005; 44: 679-86.
8. Morris C, Gibbons E, Fitzpatrick R. A structured review of patient-reported outcome measures for men with prostate cancer. Patient-reported Outcome Measurement Group, Department of Public Health, University of Oxford 2009.

9. Schmidt S, Garin O, Pardo Y, Valderas JM, Alonso J, Rebollo P, et al. Assessing quality of life in patients with prostate cancer: a systematic and standardized comparison of available instruments. *Qual Life Res* 2014; 23: 2169-81.

10. Rnic K, Linden W, Tudor I, Pullmer R, Vodermaier A. Measuring symptoms in localized prostate cancer: a systematic review of assessment instruments. *Prostate Cancer Prostatic Dis* 2013; 16: 111-22.

11. Aaronson N, Alonso J, Burnam A, Lohr KN, Patrick DL, Perrin E, et al. Assessing health status and quality-of-life instruments: attributes and review criteria. *Qual Life Res* 2002; 11: 193-205.

12. U.S. Department of Health and Human Services Food and Drug Administration. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. *Health Qual Life Outcomes* 2006; 4: 79.

13. Reeve BB, Wyrwich KW, Wu AW, Velikova G, Terwee CB, Snyder CF, et al. ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Qual Life Res* 2013; 22: 1889-905.

14. Petrillo J, Cano SJ, McLeod LD, Coon CD. Using Classical Test Theory, Item Response Theory, and Rasch Measurement Theory to Evaluate Patient-Reported Outcome Measures: A Comparison of Worked Examples. *Value in Health* 2015; 18: 25-34.

15. Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technology Assessment* 2009; 13: 200.

16. Massof RW. Understanding Rasch and Item Response Theory Models: Applications to the Estimation and Validation of Interval Latent Trait Measures from Responses to Rating Scale Questionnaires. *Ophthalmic Epidemiology* 2011; 18: 1-19.

17. Patrick DL, Erickson P. Assessing health-related quality of life for clinical decision-making. In: Walker SR, Rosser RM, eds. *Quality of Life Assessment: Key Issues in the 1990s*. Dordrecht: Springer Netherlands, 1993: 11-63.

18. Bullinger M, Anderson R, Cella D, Aaronson N. Developing and evaluating cross-cultural instruments from minimum requirements to optimal models. *Quality of Life Research* 1993; 2: 451-9.

19. WHOQOL Group. Study protocol for the World Health Organization project to develop a Quality of Life assessment instrument (WHOQOL). *Quality of Life Research* 1993; 2: 153-9.

20. Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life: A conceptual model of patient outcomes. *JAMA* 1995; 273: 59-65.

21. University of Oxford Patient Reported Outcomes Measurement Group [May 2015]. Available from: <http://phi.uhce.ox.ac.uk>.

22. Smith SC, Lamping DL, Banerjee S, Harwood R, Foley B, Smith P, et al. Measurement of health-related quality of life for people with dementia: development of a new instrument (DEMQOL) and an evaluation of current methodology. *Health Technol Assess* 2005; 9: 1-93, iii-iv.

23. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. *Quality of Life Research* 2002; 11: 193-205.

24. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual Life Res* 2012; 21: 651-7.
25. Hamoen EH, De Rooij M, Witjes JA, Barentsz JO, Rovers MM. Measuring health-related quality of life in men with prostate cancer: A systematic review of the most used questionnaires and their validity. *Urol Oncol* 2015; 33: 69.e19-28.
26. Martin NE, Massey L, Stowell C, Bangma C, Briganti A, Bill-Axelson A, et al. Defining a standard set of patient-centered outcomes for men with localized prostate cancer. *Eur Urol* 2015; 67: 460-7.
27. Szymanski KM, Wei JT, Dunn RL, Sanda MG. Development and validation of an abbreviated version of the expanded prostate cancer index composite instrument for measuring health-related quality of life among prostate cancer survivors. *Urology* 2010; 76: 1245-50.
28. Hedgepeth RC, Labo J, Zhang L, Wood DP, Jr. Expanded Prostate Cancer Index Composite versus Incontinence Symptom Index and Sexual Health Inventory for Men to measure functional outcomes after prostatectomy. *J Urol* 2009; 182: 221-7; discussion 7-8.
29. Wei JT, Dunn RL, Litwin MS, Sandler HM, Sanda MG. Development and validation of the expanded prostate cancer index composite (EPIC) for comprehensive assessment of health-related quality of life in men with prostate cancer. *Urology* 2000; 56: 899-905.
30. Wei JT, Dunn RL, Sandler HM, McLaughlin PW, Montie JE, Litwin MS, et al. Comprehensive comparison of health-related quality of life after contemporary therapies for localized prostate cancer. *J Clin Oncol* 2002; 20: 557-66.
31. Litwin MS, Hays RD, Fink A, et al. Quality-of-life outcomes in men treated for localized prostate cancer. *JAMA* 1995; 273: 129-35.

32. Litwin MS, Hays RD, Fink A, Ganz PA, Leake B, Brook RH. The UCLA Prostate Cancer Index: development, reliability, and validity of a health-related quality of life measure. *Med Care* 1998; 36: 1002-12.
33. Esper P, Mo F, Chodak G, Sinner M, Cella D, Pienta KJ. Measuring quality of life in men with prostate cancer using the functional assessment of cancer therapy-prostate instrument. *Urology* 1997; 50: 920-8.
34. van Andel G, Bottomley A, Fossa SD, Efficace F, Coens C, Guerif S, et al. An international field study of the EORTC QLQ-PR25: a questionnaire for assessing the health-related quality of life of patients with prostate cancer. *Eur J Cancer* 2008; 44: 2418-24.
35. Giesler RB, Miles BJ, Cowen ME, Kattan MW. Assessing quality of life in men with clinically localized prostate cancer: development of a new instrument for use in multiple settings. *Qual Life Res* 2000; 9: 645-65.
36. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: A Quality-of-Life Instrument for Use in International Clinical Trials in Oncology. *Journal of the National Cancer Institute* 1993; 85: 365-76.

Table 1. Psychometric definitions/tests and appraisal criteria, derived from Smith et al 2005 [22].

Psychometric property	Definition/test	Criteria for acceptability
1. Item analysis/reduction	Identification of items for possible elimination owing to weak psychometric performance; assessed on the basis of (1) unrotated principal component factor analysis to determine whether all items are measuring a single factor; and (2) item analyses for all items.	<p>Principal component factor analysis: All items should load on the first unrotated factor >0.30.</p> <p>Item analyses (applied to all items): Missing data <5%. No item redundancy (inter-item correlations ≤ 0.75). Item–total correlations >0.25 Maximum endorsement frequencies $\leq 80\%$ (i.e. the proportion of respondents who endorse each response category), including floor/ceiling effects <80% (i.e. response categories with high endorsement rates at the bottom/top ends of the scale, respectively). Aggregate adjacent endorsement frequencies $\geq 10\%$. Missing data for summary scores <5%. Even distribution of endorsement frequencies across response categories. Floor/ceiling effects for summary scores <10%.</p>
2. Acceptability	The quality of data; assessed by completeness of data and score distributions.	<p>Cronbach’s alphas for summary scores ≥ 0.70. Item–total correlations ≥ 0.20. Test–retest reliability correlations for summary scores ≥ 0.70. Test–retest reliability correlations for summary scores ≥ 0.70.</p>
3. Reliability		
3.1 Internal consistency	The extent to which items comprising a scale measure the same construct (e.g. homogeneity of the scale); assessed by Cronbach’s alpha and item–total correlations.	
3.2 Test-retest reliability	The stability of a measuring instrument; assessed by administering the instrument to respondents on two different occasions and examining the correlation between test and retest scores.	
4. Validity		
4.1 Content validity	The extent to which the content of a scale is representative of the conceptual domain it is intended to cover; assessed qualitatively during the questionnaire development stage through pre-testing with patients, expert opinion and literature review.	Qualitative evidence from pre-testing with patients, expert opinion and literature review that items in the scale are representative of the construct being measured.
4.2 Criterion-related validity		
4.2.1 Concurrent validity	Evidence that the scale predicts a gold-standard criterion that is measured at the same time; assessed on the basis of correlations between the scale and the criterion measure.	High correlation between the scale and the criterion measure.
4.2.2 Predictive validity	Evidence that the scale predicts a gold-standard criterion that is measured in the future; assessed on the basis of	High correlation between the scale and the criterion measure.

	correlations between the scale and the criterion measure.	
4.3 Construct validity		
4.3.1 Within-scale analyses	Evidence that a single entity (construct) is being measured and that items can be combined to form a summary score; assessed on the basis of evidence of good internal consistency and correlations between scale scores (which purport to measure related aspects of the construct).	Internal consistency (Cronbach's alpha) \geq 0.70. Moderate to high correlations between scale scores.
4.3.2 Analyses against external criteria		
4.3.2.1 Convergent validity	Evidence that the scale is correlated with other instruments measuring the same or similar constructs; assessed on the basis of correlations between the instrument and other similar instruments.	Correlations are expected to vary according to the degree of similarity between the constructs that are being measured by each instrument. Specific hypotheses are formulated and predictions tested on the basis of correlations.
4.3.2.2 Discriminant validity	Evidence that the scale is not correlated with instruments measuring different constructs; assessed on the basis of correlations with instruments measuring different constructs.	Low correlations between the instrument and instruments measuring different constructs.
4.3.2.3 Known groups differences	The ability of a scale to differentiate known groups; assessed by comparing scores for subgroups who are expected to differ on the construct being measured.	Significant differences between known groups or difference of expected magnitude.
4.3.2.4 Hypothesis testing	The extent to which the scale confirms pre-defined hypotheses regarding expected associations or lack of association with external factors, such as patient characteristics.	Significant moderate to high correlations, or significant associations in the expected direction. Expected lack of association confirmed.
5. Responsiveness	The ability of a scale to detect clinically important change over time; assessed by comparing scores before and after an intervention of known efficacy (on the basis of various methods including t-tests, effect sizes, standardised response means, or responsiveness statistics).	Significant differences between known groups or difference of expected magnitude.

Table 2. Description of prostate cancer specific PROs

PRO	Authors	Reviews recommending PROs	Domains (no. items)	Response options	Scoring	Target population	Recall period	Administration / completion time
STAR (Symptom Tracking and Reporting)	Vickers et al 2010 [3]	N/A	4 domains (15 items): Sexual function (6 items), urinary function (5 items), bowel function (2 items) and overall quality of life (1 item).	3- to 6-point Likert scales, and one 11-point Likert scale.	Item scores summed for the urinary and sexual function scales to give domain scores. Domain scores can also be transformed to a 0-100 scale.	Men treated for early stage prostate cancer with radical prostatectomy.	Last 4 weeks.	Self-administered. Time to complete unknown.
EPIC-26 (Expanded Prostate Index Composite-26)	Szymanski et al 2010 [27] Hedgepeth et al 2009 [28]	Schmidt et al 2014 [9] Rnic et al 2013 [10] Martin et al 2015 [26]	4 domains (26 items): Urinary domain with incontinence subscale (4), irritation/obstruction subscale (4), overall urinary bother item (1); bowel domain (6)' sexual domain (6); and vitality/hormonal domain (5).	4- and 5-point Likert scales.	Item scores summed for each scale and linearly transformed to 0-100 scale. Higher scores = better QoL.	Men treated for early stage prostate cancer with brachytherapy / external beam radiation / radical prostatectomy ± adjuvant hormonal therapy.	Last 4 weeks.	Self-administered in 10 minutes.
EPIC-50 (Expanded Prostate Index Composite)	Wei et al 2000 [29] Wei et al 2002 [30]	Schmidt et al 2014 [9] Morris et al 2009 [8]	4 domains (50 items): urinary (12) - function (5) and bother (7); bowel (14) - function (7) and bother (7); sexual (13) - function (9) and bother (4); hormonal (11) - function (5) and bother (6).	3- to 5-point Likert scales.	Item scores summed for each domain and linearly transformed to 0-100 scale. Higher scores =	Men treated for localised prostate cancer with brachytherapy / external beam radiation / radical prostatect	Last 4 weeks.	Self-administered in 15-20 minutes.

PRO	Author s	Reviews recommending PROs	Domains (no. items)	Response options	Scoring	Target population	Recall period	Administration / completion time
UCLA-PCI (University of California-Los Angeles Prostate Cancer Index)	Litwin et al 1995 [31] Litwin et al 1998 [32]	Rnic et al 2013 [10] Morris et al 2009 [8] Hamoen et al 2015 [25]	6 domains (20 items): urinary function (5), sexual function (8), bowel function (4), urinary bother (1), sexual bother (1), bowel bother (1).	3- to 5-point Likert scales.	Item scores summed for the function scales and all domain scores linearly transformed to 0-100 scales. Higher scores = better QoL.	Men treated for localised early stage prostate cancer with surgery, radiotherapy or watchful waiting.	Last 4 weeks.	Self-administered in 8-10 minutes.
FACT-P (Functional Assessment of Cancer Therapy – Prostate)	Esper et al 1997 [33]	Morris et al 2009 [8] Hamoen et al 2015 [25]	6 domains (39 items). 1 Prostate Cancer Subscale (PCS) titled 'additional concerns' (12 items). PCS 12 items: weight loss, appetite, pain bother, pain, pain activity limitation, comfort, masculine self-perception, bowel movement, difficulty urinating, urinating frequency, urinating activity limitation, erection.	5-point Likert scales.	Item scores are added to give summary score for each domain and a total FACT-P score. Physical and functional domain scores and the prostate-specific score produce the Treatment Outcome	Men with localised or metastatic prostate cancer.	Past 7 days.	Self-administered in 8-10 minutes.

PRO	Authors	Reviews recommending PROs	Domains (no. items)	Response options	Scoring	Target population	Recall period	Administration / completion time
			5 generic FACT-G domains (27 items): physical well-being (7), social/family well-being (7), emotional well-being (6), functional well-being (7).		e Index (TOI). Higher scores = better QoL. PSC score range=0-48.			
QLQ-PR25 (EORTC Prostate-specific module)	van Andel et al 2008 [34]	Morris et al 2009 [8]	6 domains (25 items): urinary symptoms (8), incontinence aid (1), bowel symptoms (4), hormonal treatment related symptoms (6), sexual active (2), sexual functioning (4).	4-point Likert scales.	Item and scale scores transformed to a 0-100 scale. Higher scores = worsening symptoms or better functioning.	Men with early or advanced localised prostate cancer.	Past week and last 4 weeks.	Self-administered in 5-10 minutes.
PC-QoL (Prostate Cancer – Quality of Life)	Giesler et al 2000 [35]	Schmidt et al 2014 [9] Rnic et al 2013 [10]	10 domains (52 items): Urinary function (5), role activity limitations (5), and bother (4); sexual function (7), role activity limitations (5), and bother (6); bowel function (7), role activity limitations (5) and bother (4); cancer worry (4).	3- to 7-point Likert scales.	Item scores summed and linearly transformed to 0-100 score range for each domain. Higher scores = better QoL.	Men with clinically localised prostate cancer treated with radical prostatectomy, radiation or watchful waiting.	Past 4 weeks.	Self-administered in 15 minutes.

Table 3. Appraisal of psychometric evidence for prostate cancer-specific PROs.

	STAR [3]	EPIC-26 [27,28]	EPIC-50 [29,30]	UCLA-PCI [31,32]	PCS (FACT-P) [33]	EORTC QLQ-PR25 [34]	PC- QoL [35]
1. Item analysis/reduction	+	++	0	++	0	0	++
2. Acceptability	+	+	++	+	+	+	+
3. Reliability							
3.1 Internal consistency	+	+++	++	++	+	++	+
3.2 Test-retest reliability	0	++	++	++	0	0	++
4. Validity							
4.1 Content validity	++	+	++	++	+	++	+++
4.2 Criterion-related validity							
4.2.1 Concurrent validity	0	+	0	0	0	0	0
4.2.2 Predictive validity	+	0	0	0	0	0	0
4.3 Construct validity							
4.3.1 Within-scale analyses	++	++	++	++	0	++	++
4.3.2 Analyses against external criteria							
4.3.2.1 Convergent validity	0	++	++	++	0	0	++
4.3.2.2 Discriminant validity	0	++	0	+	0	+	0
4.3.2.3 Known groups differences	++	0	+++	++	+	+++	++
4.3.2.4 Hypothesis testing	++	0	0	+	0	0	0
5. Responsiveness	0	0	0	0	+	+	0
6. Modern test theory	0	0	0	0	0	0	0

0: not reported or no evidence in favour; +: limited evidence in favour; ++: some acceptable evidence in favour, but some aspects fail criteria or not reported; +++: acceptable evidence in favour.







