Evaluating patient-reported outcome measures (PROMs) for bladder cancer: a systematic review using the COConsensus-based Standards for the selection of health Measurement Instruments (COSMIN) checklist

Samantha J. Mason*, James W.F. Catto†, Amy Downing*, Sarah E. Bottomley†, Adam W. Glaser* and Penny Wright*

*Leeds Institute of Cancer and Pathology, University of Leeds, Leeds, and †Academic Urology Unit, University of Sheffield, Sheffield, UK

Patient-reported outcome measures (PROMs) are important tools used to understand patient-focused outcomes from care. Various PROMs have been developed for patients with bladder cancer (BC), although the disease’s heterogeneity makes selection difficult. Accurate measurement of health-related quality of life (HRQL) can only be achieved if the PROM chosen is ‘fit for purpose’ (i.e. psychometrically sound). Systematic reviews of psychometric properties are useful for selecting the best PROM for a specific purpose.

The COConsensus-based Standards for the selection of health Measurement Instruments (COSMIN) developed a checklist to improve the selection of health measurement instruments as part of a review process. Our aims were to undertake a systematic review, using the COSMIN criteria, to assess the quality of studies that report the psychometric properties of PROMs used with people with BC and determine the psychometric quality of these PROMs. An electronic search of seven databases including PubMed, MEDLINE and EMBASE (PROSPERO reference CRD42016051974) was undertaken to identify English language publications, published between January 1990 and September 2017 that evaluated psychometric properties of PROMs used in BC research. Two researchers independently screened abstracts and selected full-text papers. Studies were rated on methodological quality using the COSMIN checklist. Overall, 4663 records were screened and 23 studies, reporting outcomes in 3568 patients, were evaluated using the COSMIN checklist. Most PROMs had limited information reported about their psychometric properties. Studies reporting on the Bladder Cancer Index (BCI) and Functional Assessment of Cancer Therapy Vanderbilt Cystectomy Index (FACT-VCI) provided the most detail and these PROMs could be evaluated on the most COSMIN properties. Based on the available evidence, no existing PROM stands out as the most appropriate to measure HRQL in BC populations. This is due to two factors; (i) the heterogeneity of BC and its treatments (ii) no PROM was evaluated on all COSMIN measurement properties due to a lack of validation studies. We suggest future evaluation of generic, cancer generic and BC-specific PROMs to better understand their application with BC populations and propose strategies to help clinicians and researchers.

Keywords
bladder cancer, COSMIN, patient reported outcome measures, psychometric properties, systematic review

Introduction
A key focus in evaluation of treatments is to have an accurate way of measuring outcomes. Survival and time to cancer progression are common primary outcomes in oncology. However, there is increased recognition of the importance of measuring what were previously regarded as softer outcomes, such as health-related quality of life (HRQL) using patient-reported outcome measures (PROMs) [1]. The number of PROMs available has grown exponentially and now there are numerous questionnaires available for assessing a multitude of domains in patients with cancer [2]. This makes it difficult to decide which PROM to use in each population. When deciding, there are a number of issues to consider: does the PROM measure what it is meant to be measuring (validity), does it do so the same way each time (reliability), and does it...
detect real differences or changes (sensitivity). Systematic reviews of studies that report on these measurement properties (also known as psychometric) are useful for selecting the best PROMs for a specific purpose [3]. A critical appraisal checklist to improve the selection of health-measurement instruments has been recommended: COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN). The checklist provides clear assessment criteria and standards, so that the methodological quality of studies that report and evaluate the psychometric properties of PROMs can be assessed as part of a review process [3,4]. This methodology is preferred over traditional evidence-based reviews, as the focus is not the evaluation of data provided by, or outcomes of PROMs, but the methodological quality of studies that report the psychometric properties of PROMs.

Bladder cancer (BC) is the ninth most common cancer worldwide and one of the most expensive malignancies to manage [5,6]. Treatment of and morbidity from non-muscle-invasive BC (NMIBC) and muscle-invasive BC (MIBC) markedly differ [7,8], and contribute to differences in patient HRQL [9]. Treatment choice will depend on the stage of the cancer, recommendations by clinicians, and patient preferences. This should be informed not only by survival rates but also HRQL outcomes. A review and meta-analysis of HRQL outcomes after radical cystectomy, where a variety of generic, cancer and BC-specific PROMs measuring HRQL were employed, found mostly low-powered studies finding similarities in HRQL between different types of diversions [10]. Comparisons with the general population showed not only poorer urinary and sexual functioning in patients after surgery, evidence was mixed about longer term outcomes [10]. HRQL in patients with NMIBC has been less well researched, possibly leading to the impression that there are few differences between the HRQL of patients with NMIBC and the general population [8]. However, a recent study developed a conceptual framework for patient-reported outcomes (PROs) in NMIBC derived from the literature, patients, and clinicians. This identified concerns about symptoms, treatment side-effects, functional problems, and experiences of care. Some of these were related to more contemporary treatments and were not included in current PROMs [11].

The heterogeneity of BC and its treatments make choosing a PROM to assess HRQL challenging [9]. Accurate measurement of HRQL in BC can only be achieved if the PROM chosen is ‘fit for purpose’ (i.e. psychometrically sound) and applied to the correct population. If the PROMs used are not ‘fit for purpose’ or are inappropriate for the patient group being studied, optimum levels of useful and informative HRQL data will not be gained from research. At worst, the data reported may be misleading or unhelpful. To date, a systematic review has not been undertaken to establish the psychometric properties of PROMs used in BC. Here we undertake a systematic review, using the COSMIN criteria. This will assess the quality of studies that report the psychometric properties of PROMs used with people with BC, determine the psychometric quality of these PROMs and identify the most promising generic, cancer-generic and BC-specific PROMs.

**Materials and Methods**

Methods were informed by the University of York, Centre for Reviews and Dissemination guidance for undertaking systematic reviews and were published on their international database of prospectively registered systematic reviews, PROSPERO (reference CRD42016051974). The COSMIN approach was employed [3,4]. Guidance can be found at the COSMIN website (http://www.cosmin.nl/). The checklist has been used in other systematic reviews of oncological HRQL instruments [12,13] and was suggested in a non-systematic review of BC HRQL research as a way to evaluate PROMs [9]. For the purposes of reporting, instruments/measures to record HRQL will be referred to as PROMs. All stages of the process following the original electronic search were undertaken by two of the research team (S.J.M., P.W.), working independently of each other and then comparing outcomes at each stage of the process. Disagreements were resolved by discussion.

**Search Strategy**

An electronic search of databases was carried out to identify publications evaluating psychometric properties of PROMs used in BC research. Searches were run in MEDLINE, EMBASE (both via OvidSP), CINAHL, PsycINFO, PubMed, The Cochrane Library, and Web of Science. Terms were agreed by the research team, appropriately modified for each database and limited to English language articles published between January 1990 and Current. Specific publication types were excluded from the search strategy, such as editorials and case reports, as per the search protocol developed by Terwee et al. [14]. A combination of Medical Subject Headings (MeSH) and free-text terms was used. Three groups of terms were generated describing: (i) the population; (ii) questionnaires, surveys and PROMs; and (iii) psychometric properties. Terms within each group were combined with the Boolean operator ‘AND’. Searches were run in November 2016, with an updated search in September 2017. The updated search included the names of PROMs found from the initial search. Reference lists of pertinent review articles identified in the literature search were checked for relevant articles, as was conference proceedings from the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) and International Society for Quality Of Life.
Research (ISOQOL) [14]. See Data S1 for an example of one database search strategy.

**Study Selection Criteria**

Inclusion criteria were English language, original studies that assessed patients with BC using a questionnaire or PROM to measure HRQL, where the study was a validation study or evaluated one or more psychometric properties of the questionnaire or PROM. Excluded were studies where the patient cohort included individuals aged <18 years, where patients had bladder problems but not BC, where patients were diagnosed with another cancer (not BC), where the PROM or questionnaire was administered by interview or by proxy, and where the PROM was a clinician-assessed instrument. Review articles, meeting abstracts, interviews, conference abstracts, editorials, and commentaries were also excluded.

Pilot testing of these criteria was undertaken to check for consensus. Testing consisted of two exercises where the reviewers (S.J.M., P.W.) independently used the inclusion and exclusion criteria to select or reject a subsample of titles and abstracts (200 in total). Following the first exercise, results were compared and interpretation of the criteria was discussed. Discrepancies were resolved through discussion. Discussions focussed on the different types of urinary diversion that may be included in BC HRQL research. Consensus was reached following the second phase of pilot testing.

Full-text papers were retrieved for titles and abstracts that either appeared to meet inclusion criteria, or where uncertainty existed. These papers were further scrutinised independently by both reviewers to identify the final list included in the review.

**Data Extraction**

The following data were extracted from each paper: PROM(s) used, constructs measured, content of the PROM and number of items/domains, psychometric information, administration method, study setting, study population, number of patients, patient demographics, response rate, country, and language. Data required to complete the COSMIN checklist assessment were also extracted.

**Appraising Methodological Quality**

The 4-point COSMIN checklist was used to evaluate the methodological quality of studies, this evaluation is important as low-quality studies are considered to have a high risk of biased results. The checklist consists of nine measurement properties, each with their own quality criteria, which form three domains (‘reliability’, ‘validity’, ‘responsiveness’). Three measurement properties are part of the reliability domain: ‘internal consistency’, ‘reliability’, and ‘measurement error’. Five measurement properties comprise the validity domain: ‘content validity’, ‘structural validity’, ‘hypothesis testing’, ‘cross-cultural validity’, and ‘criterion validity’. Responsiveness is a separate domain. An explanation of what each of the measurement properties are and COSMIN standards for reporting each measurement property is provided in Table 1.

Each eligible study was rated for measurement properties as ‘excellent’, ‘good’, ‘fair’ or ‘poor’. Checklist criteria are used to assess how well studies report each measurement property and whether they adhere to the COSMIN standards, e.g., providing evidence of adequate sample size, a priori hypotheses or how missing values were managed. An overall score is determined by taking the lowest rating gained on any of the checklist criteria for the evaluated measurement property: ‘the worst score counts’ [4].

**Reporting of Psychometric Results**

The psychometric results reported in the studies were described and categorised into the nine COSMIN measurement properties. Quality criteria proposed by Terwee et al. [15] for health-status questionnaires was used to determine whether the results for each measurement property were ‘positive’, ‘negative’ or ‘indeterminate’. An example of these criteria is that if a study reported a Cronbach’s $\alpha$ of $<0.70$ for a PROM, the internal consistency for that study would be considered a negative result.

**Levels of Evidence Appraisal**

A levels-of-evidence appraisal was undertaken to determine the overall quality of each measurement property, established in the different studies. The appraisal produced a final rating for each PROM for each of the measurement properties. All available information was synthesised, combining the results of the different studies for each PROM. PROMs were rated based on COSMIN checklist scores (reflecting the methodological quality of the studies), reported psychometric evidence, and quality of the evidence (whether results were positive, negative or indeterminate), the consistency of results between studies, and level of evidence. The levels of evidence rating could be ‘strong’ (+++ or +++), ‘moderate’ (++ or ++), ‘limited’ (+ or –), ‘conflicting’ (+/−), or ‘unknown’ (?). For example, if two studies reported a Cronbach’s $\alpha$ of $<0.70$ for a PROM, but had both scored ‘Good’ on the COSMIN checklist for internal consistency, the rating would be ‘+++’, meaning there is strong evidence (multiple studies of good methodological quality) for low levels of internal consistency. However, if there was one study reporting a Cronbach’s $\alpha$ of $>0.70$ which scored ‘Good’ on the COSMIN checklist, the rating of that PROM for internal consistency would be ‘++’, meaning there is moderate evidence in a study of good methodological quality. When there are only studies of poor methodological quality, an unknown rating is given. Levels-of-evidence criteria are presented in Table 2 [14].
<table>
<thead>
<tr>
<th>Measurement property</th>
<th>COSMIN standards for reporting the measurement property</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content validity</strong></td>
<td>Evidence should be presented of an assessment concerning the relevance and scope of the items. Does the questionnaire include items relevant to the underlying outcome (construct) of interest? In this instance, the construct is HRQL in bladder cancer patients. The domain of interest is typically specified. Often, HCQ questionnaires comprise a number of scales which represent different constructs of HRQL. Internal consistency is closely related to structural validity. Internal consistency is measured by looking at the correlation between the items within a scale and the overall scale score if that specific item was excluded. Cronbach’s α or Kuder–Richardson Formula 20 (KR-20) are used. It is measured is by looking at the correlation between the items within a scale and examining the retest reliability of the total score. These tests assess how well items fit the scale (unidimensionality) and whether they should be excluded. Tests of internal consistency include factor analysis and Item Response Theory (IRT), Rasch analysis should be described including estimations for parameters of the model. Following initial Factor analysis to check scale fit the scale (unidimensionality), Cronbach’s α ≥ 0.70 or KR-20 should be reported. Items within the same scale or domain should be moderately correlated with each other, all contributing in a different way to the overall scale score. The items within a scale should be related to each other, all contributing in a different way to the overall scale score. The scale should be reliable when it is measured twice by patients whose health is stable.</td>
</tr>
<tr>
<td><strong>Structural validity</strong></td>
<td>Evidence should be presented of an assessment concerning item relevance and scope. Does the questionnaire include items relevant to the underlying outcome (construct) of interest? In this instance, the construct is HRQL in bladder cancer patients. The domain of interest is typically specified. Often, HCQ questionnaires comprise a number of scales which represent different constructs of HRQL. Internal consistency is closely related to structural validity. Internal consistency is measured by looking at the correlation between the items within a scale and the overall scale score if that specific item was excluded. Cronbach’s α or Kuder–Richardson Formula 20 (KR-20) are used. It is measured is by looking at the correlation between the items within a scale and examining the retest reliability of the total score. These tests assess how well items fit the scale (unidimensionality) and whether they should be excluded. Tests of internal consistency include factor analysis and Item Response Theory (IRT), Rasch analysis should be described including estimations for parameters of the model. Following initial Factor analysis to check scale fit the scale (unidimensionality), Cronbach’s α ≥ 0.70 or KR-20 should be reported. Items within the same scale or domain should be moderately correlated with each other, all contributing in a different way to the overall scale score. The items within a scale should be related to each other, all contributing in a different way to the overall scale score. The scale should be reliable when it is measured twice by patients whose health is stable.</td>
</tr>
<tr>
<td><strong>Reliability</strong></td>
<td>Evidence should be presented of an assessment concerning item relevance and scope. Does the questionnaire include items relevant to the underlying outcome (construct) of interest? In this instance, the construct is HRQL in bladder cancer patients. The domain of interest is typically specified. Often, HCQ questionnaires comprise a number of scales which represent different constructs of HRQL. Internal consistency is closely related to structural validity. Internal consistency is measured by looking at the correlation between the items within a scale and the overall scale score if that specific item was excluded. Cronbach’s α or Kuder–Richardson Formula 20 (KR-20) are used. It is measured is by looking at the correlation between the items within a scale and examining the retest reliability of the total score. These tests assess how well items fit the scale (unidimensionality) and whether they should be excluded. Tests of internal consistency include factor analysis and Item Response Theory (IRT), Rasch analysis should be described including estimations for parameters of the model. Following initial Factor analysis to check scale fit the scale (unidimensionality), Cronbach’s α ≥ 0.70 or KR-20 should be reported. Items within the same scale or domain should be moderately correlated with each other, all contributing in a different way to the overall scale score. The items within a scale should be related to each other, all contributing in a different way to the overall scale score. The scale should be reliable when it is measured twice by patients whose health is stable.</td>
</tr>
<tr>
<td><strong>Hypothesis testing</strong></td>
<td>Evidence should be presented of an assessment concerning item relevance and scope. Does the questionnaire include items relevant to the underlying outcome (construct) of interest? In this instance, the construct is HRQL in bladder cancer patients. The domain of interest is typically specified. Often, HCQ questionnaires comprise a number of scales which represent different constructs of HRQL. Internal consistency is closely related to structural validity. Internal consistency is measured by looking at the correlation between the items within a scale and the overall scale score if that specific item was excluded. Cronbach’s α or Kuder–Richardson Formula 20 (KR-20) are used. It is measured is by looking at the correlation between the items within a scale and examining the retest reliability of the total score. These tests assess how well items fit the scale (unidimensionality) and whether they should be excluded. Tests of internal consistency include factor analysis and Item Response Theory (IRT), Rasch analysis should be described including estimations for parameters of the model. Following initial Factor analysis to check scale fit the scale (unidimensionality), Cronbach’s α ≥ 0.70 or KR-20 should be reported. Items within the same scale or domain should be moderately correlated with each other, all contributing in a different way to the overall scale score. The items within a scale should be related to each other, all contributing in a different way to the overall scale score. The scale should be reliable when it is measured twice by patients whose health is stable.</td>
</tr>
<tr>
<td><strong>Measurement error</strong></td>
<td>Evidence should be presented of an assessment concerning item relevance and scope. Does the questionnaire include items relevant to the underlying outcome (construct) of interest? In this instance, the construct is HRQL in bladder cancer patients. The domain of interest is typically specified. Often, HCQ questionnaires comprise a number of scales which represent different constructs of HRQL. Internal consistency is closely related to structural validity. Internal consistency is measured by looking at the correlation between the items within a scale and the overall scale score if that specific item was excluded. Cronbach’s α or Kuder–Richardson Formula 20 (KR-20) are used. It is measured is by looking at the correlation between the items within a scale and examining the retest reliability of the total score. These tests assess how well items fit the scale (unidimensionality) and whether they should be excluded. Tests of internal consistency include factor analysis and Item Response Theory (IRT), Rasch analysis should be described including estimations for parameters of the model. Following initial Factor analysis to check scale fit the scale (unidimensionality), Cronbach’s α ≥ 0.70 or KR-20 should be reported. Items within the same scale or domain should be moderately correlated with each other, all contributing in a different way to the overall scale score. The items within a scale should be related to each other, all contributing in a different way to the overall scale score. The scale should be reliable when it is measured twice by patients whose health is stable.</td>
</tr>
<tr>
<td><strong>Criterion validity</strong></td>
<td>Evidence should be presented of an assessment concerning item relevance and scope. Does the questionnaire include items relevant to the underlying outcome (construct) of interest? In this instance, the construct is HRQL in bladder cancer patients. The domain of interest is typically specified. Often, HCQ questionnaires comprise a number of scales which represent different constructs of HRQL. Internal consistency is closely related to structural validity. Internal consistency is measured by looking at the correlation between the items within a scale and the overall scale score if that specific item was excluded. Cronbach’s α or Kuder–Richardson Formula 20 (KR-20) are used. It is measured is by looking at the correlation between the items within a scale and examining the retest reliability of the total score. These tests assess how well items fit the scale (unidimensionality) and whether they should be excluded. Tests of internal consistency include factor analysis and Item Response Theory (IRT), Rasch analysis should be described including estimations for parameters of the model. Following initial Factor analysis to check scale fit the scale (unidimensionality), Cronbach’s α ≥ 0.70 or KR-20 should be reported. Items within the same scale or domain should be moderately correlated with each other, all contributing in a different way to the overall scale score. The items within a scale should be related to each other, all contributing in a different way to the overall scale score. The scale should be reliable when it is measured twice by patients whose health is stable.</td>
</tr>
<tr>
<td><strong>Responsiveness</strong></td>
<td>Evidence should be presented of an assessment concerning item relevance and scope. Does the questionnaire include items relevant to the underlying outcome (construct) of interest? In this instance, the construct is HRQL in bladder cancer patients. The domain of interest is typically specified. Often, HCQ questionnaires comprise a number of scales which represent different constructs of HRQL. Internal consistency is closely related to structural validity. Internal consistency is measured by looking at the correlation between the items within a scale and the overall scale score if that specific item was excluded. Cronbach’s α or Kuder–Richardson Formula 20 (KR-20) are used. It is measured is by looking at the correlation between the items within a scale and examining the retest reliability of the total score. These tests assess how well items fit the scale (unidimensionality) and whether they should be excluded. Tests of internal consistency include factor analysis and Item Response Theory (IRT), Rasch analysis should be described including estimations for parameters of the model. Following initial Factor analysis to check scale fit the scale (unidimensionality), Cronbach’s α ≥ 0.70 or KR-20 should be reported. Items within the same scale or domain should be moderately correlated with each other, all contributing in a different way to the overall scale score. The items within a scale should be related to each other, all contributing in a different way to the overall scale score. The scale should be reliable when it is measured twice by patients whose health is stable.</td>
</tr>
</tbody>
</table>

**Table 1** Description of the COSMIN checklist measurement properties and standards for reporting each measurement property.

© 2018 The Authors

BJU International published by John Wiley & Sons Ltd on behalf of BJU International
Results
The initial search produced 4663 results. After removal of duplicates and screening titles and abstracts, 280 full texts were agreed for further examination, of which 19 met the inclusion criteria. Three articles were included from hand-searching the reference sections of review articles and one article was included as a result of the updated search (Fig. 1). No results were found from hand-searching ISPOR or ISOQOL conference proceedings. Overall, 23 studies, reporting PROMs in 3568 patients, were evaluated using the COSMIN checklist. An overview of studies and PROMs are presented in Table 3 [16–38] and Table 4 [35–38], respectively.

We identified three generic PROMs (EuroQoL five Dimensions [EQ-5D] [16,17], 36-item short-form health survey [SF-36] [16], World Health Organisation Quality of Life [WHOQOL-BREF] [18]), one cancer-generic PROM (European Organisation for Research and Treatment of Cancer quality-of-life, 30 item core questionnaire [EORTC QLQ-C30] [17,19,20]), seven BC-specific PROMs including two for all patients with BC (Bladder Cancer Index [BCI] [16,21–25], Functional Assessment of Cancer Therapy-Bladder [FACT-B] [16,26–28]), three for MIBC (FACT Vanderbilt Cystectomy Index [FACT-VCI] [21,29–31], EORTC QLQ-BLM30, EORTC QLQ-BLM30 and EORTC

Table 2 Levels of evidence for the quality of the measurement properties for PROMs, taken from Terwee et al. [14].

<table>
<thead>
<tr>
<th>Level</th>
<th>Rating*</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong</td>
<td>+++ or −−</td>
<td>Consistent findings in multiple studies of good methodological quality OR in one study of excellent methodological quality</td>
</tr>
<tr>
<td>Moderate</td>
<td>++ or −−</td>
<td>Consistent findings in multiple studies of fair methodological quality OR in one study of good methodological quality</td>
</tr>
<tr>
<td>Limited</td>
<td>+ or −−</td>
<td>One study of fair methodological quality</td>
</tr>
<tr>
<td>Conflicting</td>
<td>+/−</td>
<td>Conflicting findings</td>
</tr>
<tr>
<td>Unknown</td>
<td>?</td>
<td>Only studies of poor methodological quality</td>
</tr>
</tbody>
</table>

*Positive rating, ‘+’; Indeterminate rating, ‘?’; Negative rating, ‘−’.

Fig. 1 Flow chart showing identification and selection of eligible articles.
Table 3 Overview of studies included in the review.

<table>
<thead>
<tr>
<th>(a) MIIBC population</th>
<th>Setting</th>
<th>Cohort</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[17] Mak et al., 2016</td>
<td>Postal study in USA using English language EQ-5D-5L (version not reported) and EORTC QLQ-BLM30</td>
<td>Patients with non-metastatic MIIBC with radical cystectomy ≥2 years disease free. 226 eligible, 177 returned questionnaires, 25% of 109 radical cystectomy patients (median age 73 years) and 22% of 64 trimodality therapy (median age 76 years), 100% female</td>
<td></td>
</tr>
<tr>
<td>[18] Liu et al., 2016</td>
<td>Recruited from Department of Urology, West China Hospital. Completed Chinese language version of WHOQOL-BREF</td>
<td>MIIBC with ileal conduit diversion with a postoperative period of 1 month to 2 years. 188 questionnaires returned, 66% aged &gt;60 years, 19% female</td>
<td></td>
</tr>
<tr>
<td>[21] Moncrief et al., 2017</td>
<td>Postal study in USA using English language BCI and FACT-VCI</td>
<td>64 patients with MIIBC with ileal conduit diversion (median age 72 years) or neobladder (median age 63 years), 25% female</td>
<td></td>
</tr>
<tr>
<td>[29] Anderson et al., 2012</td>
<td>Recruited from Vanderbilt University and University of Chicago, USA. Completed English language FACT-VCI</td>
<td>171 patients with MIIBC with radical cystectomy and urinary diversion, median (so) age 67 (10) years, 39% female</td>
<td></td>
</tr>
<tr>
<td>[30] Cookson et al., 2003</td>
<td>Postal study in USA using English language FACT-VCI</td>
<td>40 patients with MIIBC with ileal conduit or neobladder, mean (range) age 67.5 (42-87) years, 17% female</td>
<td></td>
</tr>
<tr>
<td>[31] Stenzelius et al., 2016</td>
<td>Recruited ≥1 year after surgery from Skae University Hospital, Sweden. Completed Swedish language version of FACT-VCI, which had been translated as part of the study</td>
<td>63 patients with MIIBC with urinary diversion, mean (so) age 69.8 (9.1) years, 18% female</td>
<td></td>
</tr>
<tr>
<td>[32] Siracusano et al., 2014</td>
<td>Recruited from five Italian University clinics. Completed Italian language IONB-PRO</td>
<td>171 patients with MIIBC with orthotopic neobladder, mean (so) age 64.3 (9.4) years, 9% female</td>
<td></td>
</tr>
<tr>
<td>[35] Caffo et al., 1996</td>
<td>Postal study in Italy using Italian language ad hoc PROM</td>
<td>59 patients with T2–T3 MIIBC; 30 cystectomy [median (range) age 71 (49–84) years, 13% female] and 29 conservative treatment [median (range) age 72 (40–86) years, 21% female</td>
<td></td>
</tr>
<tr>
<td>[36] Bjerre et al., 1995</td>
<td>Postal study and recruitment from outpatient clinics in Denmark using Danish language ad hoc PROM</td>
<td>67 non-malignant patients with MIIBC with urinary diversion. Excluded from study if aged ≥80 years, median (range) age 68.2 (50.8–75.7) years, 0% female</td>
<td></td>
</tr>
<tr>
<td>[37] Hart et al., 1999</td>
<td>Postal study in USA using English language ad hoc PROM</td>
<td>224 patients with MIIBC with cystectomy and either ileal conduit [mean (range) age 76.2 (58–91) years], Koch pouch [mean (range) age 70.6 (49–97) years] or urethral diversion [mean (range) age 67.3 (37–86) years], 24% female</td>
<td></td>
</tr>
<tr>
<td>[38] Henningsboh et al., 2002</td>
<td>Postal study in Denmark using a Swedish to Danish translation of ad hoc PROM</td>
<td>106 patients with MIIBC, 33% high risk, 57% aged ≥60 years, 23% female</td>
<td></td>
</tr>
<tr>
<td>[39] Stenzelius et al., 2016</td>
<td>Postal study in Denmark using a Swedish to Danish translation of ad hoc PROM</td>
<td>9 patients with MIIBC with urinary diversion (mean age 64 years) and a control group (mean age 65 years), 10% of controls and patients female</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(b) NMIBC population</th>
<th>Setting</th>
<th>Cohort</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[19] Blazeby et al., 2014</td>
<td>Recruited from Bladder COX-2 Inhibition Trial (BOXIT). Completed English language EORTC QLQ-C30 and EORTC QLQ-NMIBC24 before treatment in UK clinic at 2, 3, 6, and 12 months</td>
<td>433 patients with NMIBC; 74.6% high risk, Ta 167 (41%), T1 167 (41%), Tis 45 (11%), Ta/Tis 17 (4%) and T1/Tis 14 (3%). Mean (so) age 66.7 (9.3) years, 21% female</td>
<td></td>
</tr>
<tr>
<td>[20] Wei et al., 2014</td>
<td>Recruited from The People’s Hospital of Guangxi Zhuang Autonomous Region, China. Completed Chinese language EORTC QLQ-C30 before treatment and 6 weeks after treatment</td>
<td>106 patients with NMIBC, 33% high risk, 57% aged ≥60 years, 23% female</td>
<td></td>
</tr>
<tr>
<td>[34] Mogens et al., 2016</td>
<td>Postal study in Denmark 1 week after discharge using Danish language EORTC QLQ-NMIBC</td>
<td>121 patients with NMIBC; pTa 68 (56%), pT1a 7 (6%), carcinoma in situ 9 (7%), mean (range) age 71 (41–96) years, 31% female</td>
<td></td>
</tr>
<tr>
<td>[33] Abigail-Pedraza et al., 2016</td>
<td>Primary care in Spain using Spanish language CAIVACVENMI</td>
<td>180 patients with NMIBC, age and percentage of females not reported</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(c) All patients with BC</th>
<th>Setting</th>
<th>Cohort</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>[22,23] Gilbert et al., 2007; 2010</td>
<td>Postal study in USA using English language BCI</td>
<td>315 patients; Ta, Tis, T1 166 (53%), T2–T4 119 (38%), Unknown 30 (9%), median (range) age 69 (41–89) years, 18% female</td>
<td></td>
</tr>
<tr>
<td>[24] Heyes et al., 2016</td>
<td>Postal and recruitment at hospitals, private clinics and support groups in Australia. Completed English language BCI</td>
<td>119 patients with NMIBC and MIBC, mean (so) age 70.7 (9.6) years, 26% female</td>
<td></td>
</tr>
<tr>
<td>[25] Schmidt et al., 2014</td>
<td>197 patients; T1: 2 (2.5%), Ta: 58 (29.4%), Tis: 5 (2.5%), T1: 102 (51.8%), T2a: 16 (8.1%), T2b: 6 (3%), T3: 3 (1.5%), T4: 2 (1%), Missing: 11 (5.6%), mean (so) age 69.3 (11) years, 13% female</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
presented in Table S1. When reported, response rates were
An overview of the psychometric properties for all PROMs is
[36x83]Table 5 presents the levels of evidence for PROMs. All
response rates or percentage of missing items. Internal
consistency was usually presented as Cronbach’s \( \alpha \), although
one study reported an item-biased method but did not
present results [36]. Cronbach’s \( \alpha \) for scales and domains
were not always presented as expected. For example, a study
evaluating the BCI combined the urinary, bowel and sexual
function items and urinary, bowel and sexual bother items;
creating two new scales for which Cronbach’s \( \alpha \) was
calculated [24] instead of reporting on the scales and domains
defined in the original BCI validation work [22,23]. Test–
retest reliability was the only type of reliability reported by
studies; with many reporting Pearson’s or Spearman’s
 correlations. Intraclass correlation (ICC) or \( \kappa \) scores were less
frequently reported when reporting test–retest reliability and
some studies did not explicitly report the test carried out.
Content validity was usually undertaken using a combination
of literature searching, a working group of clinicians and
patients, and cognitive or pilot testing of items. Structural
validity was assessed by either exploratory or confirmatory
factor analysis. Rasch analysis was less common (two studies
[32,36]). Responsiveness was reported in four studies, with
one study reporting effect sizes. Although aspects of
hypothesis testing were presented in studies (internal
relationships and correlations with other instruments) because
many findings were not hypothesised a priori, these studies
were considered ‘indeterminate’.

### Levels of Evidence

Table 6 presents the levels of evidence for PROMs. All
PROMs had at least one negative or unknown rating for
COSMIN measurement properties. Despite being the most
evaluated PROM, the BCI was rated negative for the most
measurement properties; with reliability and hypothesis
testing both rated as moderate negative.

Six PROMs received strong positive ratings for content
validity. FACT-VCI was rated as unknown because the
Cookson et al. [30] paper did not assess whether all items
were relevant for the study population.

Structural validity was evaluated in eight PROMs. IONB-PRO
and CAVICAVENMI received positive ratings. FACT-VCI
received a moderate negative rating, as factors generated from
the factor analysis of postoperative data did not explain
>50%

### Psychometric Properties

An overview of the psychometric properties for all PROMs is
presented in Table S1. When reported, response rates were
between 45% and 98%, although many studies did not report

### Methodological Quality

Table 5 presents COSMIN checklist scores, assessing
methodological quality of studies that reported COSMIN
measurement properties for PROMs. The most frequently
reported properties were internal consistency (18 studies) and
hypothesis testing (15 studies). Criterion validity was not
reported in any study due to lack of a ‘gold standard’,
according to the COSMIN definition. Studies reporting on the
BCI and FACT-VCI provided the most detail and these
PROMs could be evaluated on the most COSMIN
measurement properties (seven properties).

The best performing property was content validity, with six of
the seven reporting studies receiving a score of ‘excellent’. None
of the other properties were given this quality rating due to the
COSMIN rule of ‘worst score counts’. For example, any study
not reporting quantity of missing data or how missing values
were managed could not receive an excellent rating on any
property that included missing values as part of the assessment.
The property with the worst performance was cross-cultural
validity, where all six studies scored ‘poor’ because factor
analysis was not undertaken or the sample size was inadequate.

### Psychometric Properties

An overview of the psychometric properties for all PROMs is
presented in Table S1. When reported, response rates were
between 45% and 98%, although many studies did not report

### Methodological Quality

Table 5 presents COSMIN checklist scores, assessing
methodological quality of studies that reported COSMIN
measurement properties for PROMs. The most frequently
reported properties were internal consistency (18 studies) and
hypothesis testing (15 studies). Criterion validity was not
reported in any study due to lack of a ‘gold standard’,
according to the COSMIN definition. Studies reporting on the
BCI and FACT-VCI provided the most detail and these
PROMs could be evaluated on the most COSMIN
measurement properties (seven properties).

The best performing property was content validity, with six of
the seven reporting studies receiving a score of ‘excellent’. None
of the other properties were given this quality rating due to the
COSMIN rule of ‘worst score counts’. For example, any study
not reporting quantity of missing data or how missing values
were managed could not receive an excellent rating on any
property that included missing values as part of the assessment.
The property with the worst performance was cross-cultural
validity, where all six studies scored ‘poor’ because factor
analysis was not undertaken or the sample size was inadequate.

### Psychometric Properties

An overview of the psychometric properties for all PROMs is
presented in Table S1. When reported, response rates were
between 45% and 98%, although many studies did not report

### Methodological Quality

Table 5 presents COSMIN checklist scores, assessing
methodological quality of studies that reported COSMIN
measurement properties for PROMs. The most frequently
reported properties were internal consistency (18 studies) and
hypothesis testing (15 studies). Criterion validity was not
reported in any study due to lack of a ‘gold standard’,
according to the COSMIN definition. Studies reporting on the
BCI and FACT-VCI provided the most detail and these
PROMs could be evaluated on the most COSMIN
measurement properties (seven properties).

The best performing property was content validity, with six of
the seven reporting studies receiving a score of ‘excellent’. None
of the other properties were given this quality rating due to the
COSMIN rule of ‘worst score counts’. For example, any study
not reporting quantity of missing data or how missing values
were managed could not receive an excellent rating on any
property that included missing values as part of the assessment.
The property with the worst performance was cross-cultural
validity, where all six studies scored ‘poor’ because factor
analysis was not undertaken or the sample size was inadequate.

### Psychometric Properties

An overview of the psychometric properties for all PROMs is
presented in Table S1. When reported, response rates were
between 45% and 98%, although many studies did not report

### Methodological Quality

Table 5 presents COSMIN checklist scores, assessing
methodological quality of studies that reported COSMIN
measurement properties for PROMs. The most frequently
reported properties were internal consistency (18 studies) and
hypothesis testing (15 studies). Criterion validity was not
reported in any study due to lack of a ‘gold standard’,
according to the COSMIN definition. Studies reporting on the
BCI and FACT-VCI provided the most detail and these
PROMs could be evaluated on the most COSMIN
measurement properties (seven properties).

The best performing property was content validity, with six of
the seven reporting studies receiving a score of ‘excellent’. None
of the other properties were given this quality rating due to the
COSMIN rule of ‘worst score counts’. For example, any study
not reporting quantity of missing data or how missing values
were managed could not receive an excellent rating on any
property that included missing values as part of the assessment.
The property with the worst performance was cross-cultural
validity, where all six studies scored ‘poor’ because factor
analysis was not undertaken or the sample size was inadequate.
Table 4 Overview of the PROMs that were evaluated.

<table>
<thead>
<tr>
<th>Name of PROM</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQ-5D</td>
<td>5 items: mobility, self-care, usual activities, pain, anxiety and depression, plus a visual analogue scale</td>
</tr>
<tr>
<td>SF-36</td>
<td>36 items overall, mental component scale and physical component scale</td>
</tr>
<tr>
<td>WHOQOL-BREF</td>
<td>26 items; one from each of the 24 facets of WHOQOL100 plus 2 items from quality of life and general health items. Scales were physical health, psychological health, social relationships, and environment</td>
</tr>
<tr>
<td>EORTC QLQ-C30</td>
<td>17 items specific to bladder cancer plus 27 item FACT-G that comprises 4 scales; functional, social/family and physical wellbeing scales (7 items), emotional wellbeing scale (6 items)</td>
</tr>
<tr>
<td>FACT-Bl</td>
<td>13 items specific to bladder cancer patients, plus 27 item FACT-G that comprises 4 scales; functional, social/family and physical wellbeing scales (7 items), emotional wellbeing scale (6 items)</td>
</tr>
<tr>
<td>FACT-VCI</td>
<td>17 items specific to radical cystectomy patients, plus 27 item FACT-G that comprises 4 scales; functional, social/family and physical wellbeing scales (7 items), emotional wellbeing scale (6 items)</td>
</tr>
<tr>
<td>IONB-PRO</td>
<td>Relational, fatigue, and emotional scales. Analysis from research produced several other scales</td>
</tr>
<tr>
<td>Caffo et al. [35]</td>
<td>Ad hoc PROM. 40 items for cystectomy version, 41 for conservative treatment version. Items about physical wellbeing, pain, bowel function, urinary function, sexual function, daily physical activities, relational and recreational activities, stoma, and general items</td>
</tr>
<tr>
<td>Bjerre et al. [36]</td>
<td>Ad hoc PROM. 211 items including items about urine leakage, diarrhoea, and skin complications</td>
</tr>
<tr>
<td>Hart et al. [37]</td>
<td>Ad hoc PROM. About 105 items on emotional distress, quality of life (global), sexual dissatisfaction, body image, urinary diversion problems, sexual function, physical symptomology, daily living activities</td>
</tr>
<tr>
<td>Henningsohn et al. [38]</td>
<td>Ad hoc PROM. 137 items for patients, 125 for controls about urinary symptoms, bowel, and sexual dysfunction</td>
</tr>
<tr>
<td>EORTC QLQ-NMIBC24</td>
<td>24 items. 6 scales: urinary symptoms/problems, urostomy problems, future perspective, bloating and flatulence, body image and sexuality, plus single item about catheter use</td>
</tr>
<tr>
<td>CAVICAVENMI</td>
<td>21 items. 5 scales: disease, self-esteem and emotional status, working life, daily life, and sex life</td>
</tr>
</tbody>
</table>

of the variance [29]. This study also presented factor analysis based on preoperative data, which was not considered appropriate as FACT-VCI is designed for use with patients who have undergone cystectomy. Five PROMs were rated unknown as the COSMIN checklist deemed the studies reporting on these PROMs to be of poor methodological quality due to inadequate sample sizes.

Internal consistency was evaluated in eleven PROMs. The BCI and FACT-VCI were rated moderate positive. Four PROMs and three ad hoc PROMs were rated unknown, as the reporting studies were deemed to be of poor methodological quality because factor analysis had not been performed or the sample size for the analysis was too small.

Three PROMs measured test–retest reliability. The BCI was rated moderate negative as some Pearson’s r figures were <0.80. CAVICAVENMI was rated unknown as the study scored poor on the COSMIN checklist for methodological quality [33]. Small sample sizes accounted for poor COSMIN checklist scores.

The IONB-PRO, the only PROM for which measurement error was reported, received an unknown rating as the study calculated the standard error measurement based on Cronbach’s α [32].

In all, 10 PROMs were evaluated for hypothesis testing. The FACT-VCI and EORTC QLQ-NMIBC24, were rated moderate positive. The BCI, FACT-Bl, EQ-5D and SF-36 received negative ratings as <75% of the a priori hypotheses were met and correlations with related constructs were lower than with unrelated ones. EORTC QLQ-C30 was rated unknown, as it was difficult to establish which constructs were similar between the PROM and comparator PROMs.

Responsiveness was evaluated in four PROMs. The BCI was rated unknown as effect sizes were reported rather than correlations or area-under-curve, as required by COSMIN. The EORTC QLQ-NMIBC24 was moderate negative, as <75% of the results were in accordance with the stated hypotheses.

Five PROMs were evaluated for cross-cultural validity and received unknown ratings.

Discussion

The aim of the present review was to provide robust information on the psychometric properties of PROMs used with BC cohorts, to aid selection of the most appropriate PROMs for HRQL assessment. In all, 23 studies were identified that reported measurement properties of 15 PROMs used with patients with BC over the last 27 years. None of the reviewed studies that reported psychometric properties of PROMs met all COSMIN criteria standards for methodological quality. Only two measures are applicable to all BC-specific groups. Therefore, clinicians and researchers will have to choose the ‘best fit’ to serve their assessment objective. This may result in using a combination of generic, cancer-generic and BC-specific PROMs that reach the highest COSMIN standards.
<table>
<thead>
<tr>
<th>Measurement properties</th>
<th>Reference</th>
<th>Content validity</th>
<th>Structural validity</th>
<th>Internal consistency</th>
<th>Reliability</th>
<th>Measurement error</th>
<th>Hypothesis testing</th>
<th>Criterion validity</th>
<th>Responsiveness</th>
<th>Cross-cultural validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generic</td>
<td>EQ-5D</td>
<td>Fair</td>
<td>Fair</td>
<td>Fair</td>
<td>Fair</td>
<td>Fair</td>
<td>Fair</td>
<td>Fair</td>
<td>Fair</td>
<td>Poor</td>
</tr>
<tr>
<td></td>
<td>SF-36</td>
<td>Fair</td>
<td>Fair</td>
<td>Fair</td>
<td>Fair</td>
<td>Fair</td>
<td>Fair</td>
<td>Fair</td>
<td>Fair</td>
<td>Poor</td>
</tr>
<tr>
<td></td>
<td>WHOQOL-BREF</td>
<td>Poor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cancer generic</td>
<td>EORTC QLQ-C30</td>
<td>Good</td>
<td>Good</td>
<td>Good</td>
<td>Fair</td>
<td>Good</td>
<td>Good</td>
<td>Fair</td>
<td>Fair</td>
<td>Good</td>
</tr>
<tr>
<td>Bladder Cancer</td>
<td>BCI</td>
<td>Excellent</td>
<td>Good</td>
<td>Good</td>
<td>Fair</td>
<td>Fair</td>
<td>Fair</td>
<td>Fair</td>
<td>Fair</td>
<td>Fair</td>
</tr>
<tr>
<td></td>
<td>[22]</td>
<td>[16]</td>
<td>[17]</td>
<td>[20]</td>
<td>[21]</td>
<td>[19]</td>
<td>[20]</td>
<td>[21]</td>
<td>[22]</td>
<td>[19]</td>
</tr>
<tr>
<td></td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Fair</td>
<td>Fair</td>
<td>Fair</td>
<td>Fair</td>
<td>Poor</td>
<td>Poor</td>
</tr>
<tr>
<td></td>
<td>FACT-BI</td>
<td>Fair</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[23]</td>
<td>[16]</td>
<td>[20]</td>
<td>[21]</td>
<td></td>
<td>[19]</td>
<td>[20]</td>
<td>[21]</td>
<td>[22]</td>
<td>[19]</td>
</tr>
<tr>
<td></td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Fair</td>
<td>Fair</td>
<td>Fair</td>
<td>Fair</td>
<td>Poor</td>
<td>Poor</td>
</tr>
<tr>
<td>MIBC</td>
<td>FACT-VCI</td>
<td>Poor</td>
<td>Good</td>
<td>Good</td>
<td>Fair</td>
<td>Poor</td>
<td>Fair</td>
<td>Fair</td>
<td>Fair</td>
<td>Fair</td>
</tr>
<tr>
<td></td>
<td>[24]</td>
<td>[30]</td>
<td>[31]</td>
<td>[21]</td>
<td></td>
<td>[32]</td>
<td>[33]</td>
<td>[34]</td>
<td>[35]</td>
<td>[36]</td>
</tr>
<tr>
<td></td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Fair</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
</tr>
<tr>
<td>IONB-PRO</td>
<td>EORTC QLQ-BLM30</td>
<td>Fair</td>
<td>Fair</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Fair</td>
<td>Poor</td>
<td>Poor</td>
</tr>
<tr>
<td></td>
<td>[32]</td>
<td>[17]</td>
<td>[19]</td>
<td>[22]</td>
<td></td>
<td>[23]</td>
<td>[24]</td>
<td>[25]</td>
<td>[26]</td>
<td>[27]</td>
</tr>
<tr>
<td></td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Fair</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
</tr>
<tr>
<td>Untified</td>
<td>EORTC QLQ-BLM30</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
</tr>
<tr>
<td></td>
<td>[37]</td>
<td>[38]</td>
<td>[39]</td>
<td>[40]</td>
<td></td>
<td>[41]</td>
<td>[42]</td>
<td>[43]</td>
<td>[44]</td>
<td>[45]</td>
</tr>
<tr>
<td></td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Fair</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
</tr>
<tr>
<td>NMIBC</td>
<td>EORTC QLQ-NMIBC24</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
</tr>
<tr>
<td></td>
<td>[19]</td>
<td>[20]</td>
<td>[21]</td>
<td>[22]</td>
<td></td>
<td>[23]</td>
<td>[24]</td>
<td>[25]</td>
<td>[26]</td>
<td>[27]</td>
</tr>
<tr>
<td></td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Fair</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
</tr>
<tr>
<td></td>
<td>CAVICAVENMI</td>
<td>Excellent</td>
<td>Good</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Fair</td>
<td>Poor</td>
</tr>
<tr>
<td></td>
<td>[35]</td>
<td>[36]</td>
<td>[37]</td>
<td>[38]</td>
<td></td>
<td>[39]</td>
<td>[40]</td>
<td>[41]</td>
<td>[42]</td>
<td>[43]</td>
</tr>
</tbody>
</table>
Three generic PROMs were identified in the review, all had limited psychometric properties reported in studies; the EQ-5D, SF-36, and WHOQOL-BREF [16–18]. All three generic PROMs are established internationally with multiple non-BC studies contributing normative and psychometric properties data. Consequently, they are accepted as being valid, reliable, sensitive, and applicable to a wide range of health problems [39]. However, they may lack sensitivity in measuring cancer and BC-specific issues, which is where more specialised PROMs, have a part to play. If being used alongside other more specific PROMs, clinicians and researchers may want to consider the following as part of their decision-making process: the scope of the generic PROM, the number of items included, ease of administration, and permissions and costs.

Cancer-generic PROMs, such as the EORTC QLQ-C30 and FACT – General (FACT-G), provide information about patient experiences of general symptoms, such as nausea, pain and fatigue, and the impact of cancer on a patient’s daily life, emotional health, and relationships. They can be used to compare the HRQL of patients with BC with other cancer populations. Although used as a stand-alone cancer-generic PROM in BC research, FACT-G was not evaluated separately in the present review due to its inclusion as part of the FACT-Bl and FACT-VCI.

As with generic PROMs, psychometric evaluation of the EORTC QLQ-C30 was not the focal point of studies included in the present review and thus very few psychometric properties were reported. During EORTC QLQ-C30 development, the PROM was psychometrically evaluated in a culturally diverse sample of patients with lung cancer [40] and by 2008, EORTC published reference data for a variety of cancer groups (but not for patients with BC) [41]. Despite this, EORTC QLQ-C30 is used in BC HRQL research.

BC-specific PROMs focus on commonly reported BC problems such as with urinary, bowel and sexual function. Findings reported using these more specific PROMs may inform BC treatment regimens, policy and patient support. In the main, BC-specific PROMs have been developed more recently than their generic and cancer-generic counterparts and therefore have not been as widely used in research. There were 11 BC-specific PROMs reviewed, of variable quality. On choosing which PROM to use, consideration should be given to how well it performed in the COSMIN review and the population of patients with BC being assessed.

If the target population includes the entire BC spectrum, then the choice of BC-specific PROM must have been designed for this purpose. Only two PROMs fit this brief, the BCI and FACT-Bl. Our present review found more support for the BCI, as it was more frequently evaluated in research and scored more favourably using COSMIN. However, the BCI is not without flaw, as evidence has suggested it may be difficult to interpret in circumstances where function and bother scores are different for

<table>
<thead>
<tr>
<th>Generic</th>
<th>EQ-5D</th>
<th>SF-36</th>
<th>WHOQOL-BREF</th>
<th>SF-36</th>
<th>SF-36</th>
<th>SF-36</th>
<th>SF-36</th>
<th>SF-36</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer-generic</td>
<td>EQ-5D</td>
<td>SF-36</td>
<td>SF-36</td>
<td>SF-36</td>
<td>SF-36</td>
<td>SF-36</td>
<td>SF-36</td>
<td></td>
</tr>
<tr>
<td>EORTC QLQ-C30</td>
<td>FACT-Bl</td>
<td>FACT-Bl</td>
<td>FACT-Bl</td>
<td>FACT-Bl</td>
<td>FACT-Bl</td>
<td>FACT-Bl</td>
<td>FACT-Bl</td>
<td></td>
</tr>
<tr>
<td>Bladder Cancer</td>
<td>BCI</td>
<td>BCI</td>
<td>BCI</td>
<td>BCI</td>
<td>BCI</td>
<td>BCI</td>
<td>BCI</td>
<td></td>
</tr>
<tr>
<td>NMIBC</td>
<td>EORTC QLQ-NMIBC24</td>
<td>EORTC QLQ-NMIBC24</td>
<td>EORTC QLQ-NMIBC24</td>
<td>EORTC QLQ-NMIBC24</td>
<td>EORTC QLQ-NMIBC24</td>
<td>EORTC QLQ-NMIBC24</td>
<td>EORTC QLQ-NMIBC24</td>
<td></td>
</tr>
<tr>
<td>MIBC</td>
<td>FACT-Bl</td>
<td>FACT-Bl</td>
<td>FACT-Bl</td>
<td>FACT-Bl</td>
<td>FACT-Bl</td>
<td>FACT-Bl</td>
<td>FACT-Bl</td>
<td></td>
</tr>
<tr>
<td>IONB-PRO</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td></td>
</tr>
<tr>
<td>EORTC QLQ-BLM30</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td></td>
</tr>
</tbody>
</table>

© 2018 The Authors
BJU International published by John Wiley & Sons Ltd on behalf of BJU International
a domain, under which circumstances researchers must choose whether HRQL should be determined by symptoms (function) or importance of symptoms to the patient (bother) [42]. Consequently, it was recommended that a generic or cancer-generic PROM be administered alongside the BCI [42]. In comparison, FACT-BI scored poorly on the COSMIN checklist for all reported properties. However, a recent overview paper of BC PROMs suggests that the FACT-BI performs better than indicated in the present review. The authors state each new Functional Assessment of Chronic Illness Therapy (FACTIT) scale undergoes an assessment of test–retest reliability, responsiveness, and convergent and divergent validity with 50 patients [42] (confirmed by personal correspondence with FACTIT, by S.J.M. on 14/09/2017). These psychometric data are neither published nor available within the public domain, meaning a complete COSMIN assessment cannot be undertaken, and therefore recommendation for use is not possible.

Seven MIBC-specific PROMs were reported. The FACT-VCI was the most evaluated in studies and had the most positive COSMIN ratings. It is used in HRQL studies comparing types of diversion, but is unsuitable to use with conservative treatment patients. Comparatively, the EORTC QLQ-BLM30 was assessed for two measurement properties only, but can compare radical and conservative treatment-related HRQL. Although only evaluated by one study, the IONB-PRO scored well using the COSMIN checklist and appears to be a viable tool to administer to MIBC patients with neobladders. Four studies reported psychometric properties for a number of unnamed, ad hoc PROMs [35–38], with no traceable evidence of further use.

Two PROMs were identified that measure HRQL in patients with NMIBC; both performing well when assessed using COSMIN. Although the case could be made to use either the EORTC QLQ-NMIBC24 or CAVICAVENMI in research, the EORTC QLQ-NMIBC24 has been evaluated more in research and can be used alongside EORTC QLQ-C30. However, should clinicians and researchers prefer to use a PROM that includes a combination of cancer-generic and BC-specific items that comprises fewer items, they may wish to use CAVICAVENMI over the EORTC QLQ-BLM30 combined with EORTC QLQ-C30. Findings are described in Table 7.

### Strengths and Limitations

Although a strength of the present research is that methodological quality was assessed using the robust COSMIN checklist, carrying out the appraisal provided challenges. Many studies were not written with COSMIN criteria in mind, meaning interpretation of which COSMIN measurement properties were being assessed was sometimes difficult. For example, Blazeby et al. [19] employed multi-trait scaling when evaluating the structural validity of the EORTC QLQ-NMIBC24. Multi-trait scaling is not referred to in the COSMIN guidance. Advice from COSMIN (S.J.M. personal communication with Terwee 08/05/2017) stated that this method was not appropriate to assess structural validity. The complexity of applying the checklist appears to be acknowledged by COSMIN as the checklist is under review [43] and their website has a frequently asked question (FAQ) section, regarding which measurement properties should be assessed when terminology other than that used by COSMIN is reported. Studies that received a negative or unknown rating, published following the publication of the COSMIN guidance, may have received a more favourable COSMIN.

### Table 7 Summary of review findings.

<table>
<thead>
<tr>
<th>Bladder cancer population</th>
<th>Best performing PROM using COSMIN</th>
<th>Further information</th>
<th>Alternatives</th>
<th>Further information</th>
</tr>
</thead>
<tbody>
<tr>
<td>All (or a variety of stages/grades)</td>
<td>BCI</td>
<td>Suggested that BCI is used alongside generic and/or cancer-generic PROMs as interpretation can be difficult</td>
<td>FACT-BI</td>
<td>FACT-BI can be used to collect BC-specific and cancer-generic HRQL information, as the PROM includes FACT-G and EORTC QLQ-BLM30 can be used to collect BC-specific and cancer-generic HRQL information, as the PROM can be used with EORTC QLQ-C30</td>
</tr>
<tr>
<td>MIBC</td>
<td>FACT-VCI</td>
<td>FACT-VCI can be used to collect BC-specific and cancer-generic HRQL information, as the PROM includes FACT-G</td>
<td>EORTC QLQ-BLM30</td>
<td>EORTC QLQ-BLM30 can be used to collect BC-specific and cancer-generic HRQL information, as the PROM can be used with EORTC QLQ-C30</td>
</tr>
<tr>
<td>NMIBC</td>
<td>Difficult to determine as CAVICAVENMI only evaluated in one study and both EORTC QLQ-NMIBC24 and CAVICAVENMI scored well</td>
<td>EORTC QLQ-NMIBC24 can be used to collect BC-specific and cancer-generic HRQL information, as the PROM can be used with EORTC QLQ-C30</td>
<td>IONB-PRO</td>
<td>IONB-PRO is only suitable to use with patients who have a neobladder</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CAVICAVENMI</td>
<td>CAVICAVENMI incorporates both BC-specific and cancer-generic questions within fewer items than the combination of EORTC QLQ-C30 and EORTC QLQ-NMIBC24</td>
</tr>
</tbody>
</table>
rating if their guidelines had been taken into account when reporting the psychometric properties of PROMs.

**Recommendations**

The emergence of new questionnaires, currently published as conference abstracts only [44] and the identification of the four ad hoc MIBC PROMs, indicate existing PROMs used in BC research may not be perceived as adequate by the research and clinical community. Compared with previous COSMIN reviews [13,45], the present review identified fewer PROM-validation studies reporting an array of psychometric data. Unlike in other cancer populations, the psychometric properties of generic and general cancer PROMs are less well understood in BC populations. There would be a benefit in pooling data from studies that have used generic, cancer-generic and BC-specific PROMs to facilitate the undertaking of a more detailed psychometric analysis. This would enable organisations such as the EORTC to determine reference values for patients with BC and provide research teams with more information when choosing PROM(s) for HRQL research.

Any future psychometric evaluation of generic, cancer-generic and BC-specific PROMs should be reported and published, ideally using the COSMIN guidelines, so that it can be determined how useful these PROMs are with their intended BC populations.

Future HRQL research should implement the recommendation from previous research that BC-specific PROMs be used alongside cancer-generic PROMs in order to gain comprehensive information. Furthermore, it is recommended that generic PROMs should also be administered alongside cancer-generic and BC-specific PROMs to provide a robust picture of HRQL.

**Acknowledgments**

The authors thank patients and clinicians who contributed to the work reviewed within this manuscript. The authors also thank Dr Caroline Terwee for providing additional guidance and correspondence regarding the interpretation of COSMIN. This work was funded by a project grant from Yorkshire Cancer Research (S385: The Yorkshire Cancer Research Bladder Cancer Patient Reported Outcomes Survey).

**Conflicts of Interest**

All the authors declare no conflict of interest with this work.

**Funding Sources**

This work was funded by Yorkshire Cancer Research (Study S385). The funder had no role in the design, analysis or collection of the data; in writing the manuscript; or in the decision to submit the manuscript for publication. We are grateful for the oversight provided by the sponsor, University of Sheffield.

**Financial Disclosure**

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

**References**

15. Terwee CB, Bot SDM, de Boer MR et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007; 60: 34–42


e-mail: s.j.mason@leeds.ac.uk

Abbreviations: BCI, Bladder Cancer Index; COSMIN, CONsensus-based Standards for the selection of health Measurement Instruments; MIBC questionnaire, EORTC QLQ comprised of 30 MIBC-specific items; EORTC QLQ-C30, European Organisation for Research and Treatment of Cancer quality of life questionnaire – 30-item core; NMIBC questionnaire, EORTC QLQ comprised of 24 NMIBC-specific items; EQ-5D, EuroQoL five Dimensions; FACT, Functional Assessment of Chronic Illness Therapy; FACT-(B)(G)(VCI), Functional Assessment of Cancer Therapy – (Bladder) (General) (Vanderbilt Cystectomy Index); HRQL, health-related quality of life; ICC, intraclass correlation; IONB-PRO,
Ileal Orthotopic Neobladder PRO; ISOQOL, International Society for Quality Of Life Research; ISPOR, International Society for Pharmacoeconomics and Outcomes Research; (MI)(NMI)BC, (muscle-invasive) (non-muscle-invasive) bladder cancer; PRO(M), patient-reported outcome (measure); SF-36, 36-item short-form health survey; WHOQOL-BREF, World Health Organisation Quality of Life.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article:

**Data S1.** Search strategy used for Ovid databases.

**Table S1.** Psychometric properties of all PROMs, categorised into the COSMIN measurement properties.