Systematic Review of Patient-Reported Outcome Measures for Cancer Immunotherapy Based on COSMIN Guidelines: Post-Print

Authors: Su Zhenzhen, Wang Yixuan, Zhang Liyan, Lian Xuemin, Liu Dan, Zhang Liyan

Date: 2025-09-01T00:00:00+00:00

Abstract

Background The precise assessment of immune-related adverse events (irAEs) in cancer patients through patient-reported outcomes (PROs) facilitates early identification of irAEs and timely development and implementation of targeted interventions, ensuring continuous treatment and favorable prognosis. Currently, commonly used patient-reported outcome measures (PROMs) for cancer immunotherapy are primarily generic instruments with poor content validity, failing to capture nearly 30% of common irAEs. Disease-specific PROMs have heterogeneous items, lack unified standards, and have not undergone systematic evaluation of measurement properties, making it difficult to select optimal assessment tools.

Objective To evaluate the measurement properties of PROMs for cancer immunotherapy and provide an evidence-based basis for healthcare professionals to accurately assess irAEs and quality of life in cancer patients.

Methods According to the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) guideline requirements, we searched CNKI, Wanfang Data Knowledge Service Platform, SinoMed, PubMed, Embase, CINAHL, and ProQuest databases from inception to December 31, 2024, including studies that evaluated at least one measurement property of PROMs for cancer immunotherapy. Literature screening and data extraction were independently conducted by two researchers. Quality assessment was performed using the Chinese version of the COSMIN risk of bias checklist and the COSMIN content validity rating system. Finally, the modified Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach for quantitative systematic reviews was used to formulate recommendation grades and recommendations for the measurement instruments.

Results A total of 9 studies were included, involving nine immunotherapy PROMs: Functional Assessment of Cancer Therapy-Immune Checkpoint Inhibitor Module (FACT-ICM), Chinese version of FACT-ICM (C-FACT-ICM), Patient-Reported Outcome Measure for Financial Toxicity (PROFFIT), Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE) lung cancer subscale, MD Anderson Symptom Inventory-Immunotherapy Early Phase Trial (MDASI-Immunotherapy EPT), Chinese version of MDASI-Immunotherapy EPT, PRO-CTCAE subset for lung cancer immunotherapy, Cancer Immunotherapy Patient Symptom Self-Report Scale, and Lung Cancer Patient Immune-Related Adverse Event Self-Report Symptom Scale. None reported cross-cultural validity, measurement error, or responsiveness. Regarding content validity, FACT-ICM and PROFFIT were rated as "adequate," while all other scales were "uncertain." Regarding internal consistency, FACT-ICM was not validated, and PROFFIT was rated as "inadequate." All scales failed to meet the criterion of having high-quality evidence that any measurement property was "inadequate (-)," resulting in Grade B recommendation for all nine tools.

Conclusion C-FACT-ICM can be tentatively recommended (Grade B recommendation level). Future studies should use this scale to measure patient-reported outcomes across various cancer immunotherapy populations to enhance its clinical applicability and utility. Overall, the methodological quality of studies on PROMs for cancer immunotherapy and the measurement properties of these tools require further validation and improvement.

Full Text

Systematic Review of Patient-Reported Outcome Measures for Cancer Immunotherapy Based on the COSMIN Guidelines

SU Zhenzhen[1], WANG Yixuan[2], ZHANG Liyan[1]*, LIAN Xuemin[3], LIU Dan[2]

[1]Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education)/Department of Gastrointestinal Oncology, Peking University Cancer Hospital & Institute, Beijing 100142, China
[2]Peking University School of Nursing, Beijing 100191, China
[3]Department of Health and Medical Care, Tianjin Medical University General Hospital, Tianjin 300052, China

Corresponding author: ZHANG Liyan, Chief Superintendent Nurse; E-mail: zl54920@bjmu.edu.cn

Abstract

Background: Accurate assessment of immune-related adverse events (irAEs) in cancer patients through patient-reported outcomes facilitates early identification of irAEs and enables timely development and implementation of targeted intervention measures, which are essential for ensuring continued treatment and favorable prognosis. Currently, commonly used patient-reported outcome measures (PROMs) for tumor immunotherapy are mainly universal scales with poor content validity. Nearly 30% of common irAEs cannot be measured, and specific PROMs items vary, with no unified standards. There is also a lack of systematic evaluation of measurement performance, making it impossible to select the best assessment tool. **Objective:** To evaluate the psychometric properties of PROMs for cancer immunotherapy, providing evidence-based support for healthcare professionals to accurately assess immune-related adverse events and quality of life in cancer patients. **Methods:** According to the Consensus-Based Standards for the Selection of Health Measurement Tools (COSMIN) guidelines, we searched the following databases: CNKI, Wanfang Data, SinoMed, PubMed, Embase, CINAHL, and ProQuest. The search period was from the establishment of the databases to December 31, 2024. Studies that included at least one evaluation of the measurement properties of PROMs related to tumor immunotherapy were included. Literature screening and data extraction were performed independently by two researchers. Quality assessment was conducted using the Chinese version of the COSMIN bias risk assessment checklist and the Chinese version of the COSMIN content validity scoring system. Finally, the modified quantitative system for evaluating evidence grading was used to form the recommended level and recommendation for the measurement tool. **Results:** A total of 9 studies were included, covering the Functional Assessment of Cancer Therapy-immune Checkpoint Inhibitor Treatment Patient-Specific Module (FACT-ICM), Chinese version of FACT-ICM (C-FACT-ICM), patient-reported outcome measures for anti-economic toxicity, patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE) lung cancer subscale, MD Anderson Symptom Inventory module specific to immunotherapy for early-phase trials (MDASI-Immunotherapy EPT), Chinese version of MDASI-Immunotherapy EPT, lung cancer immunotherapy PRO-CTCAE subset, cancer immunotherapy patient symptom self-report scale, and lung cancer patient immune-related adverse event self-report symptom scale, totaling 9 immunotherapy PROMs. None of these PROMs reported cross-cultural validity, measurement error, or responsiveness. In terms of content validity, FACT-ICM and PROFFIT were rated as "adequate," while the remaining scales were rated as "uncertain." In terms of internal consistency, FACT-ICM was not validated, and PROFFIT was rated as "inadequate." None of the scales met the criteria for "inadequate (-)" for any measurement attribute based on high-level evidence. All nine tools were assigned a B-level recommendation. **Conclusion:** The C-FACT-ICM can be temporarily recommended for use (Grade B recommendation). In the future, this scale can be used to measure patient-reported outcomes in various types of tumor immunotherapy to improve

its clinical applicability and practicality. However, overall, the methodological quality of tumor immunotherapy PROMs-related research and the measurement properties of the tools still need to be further verified and improved.

Keywords: Cancer; Immunotherapy; Patient-reported outcome; Assessment tools; Systematic review; Psychometrics; COSMIN

Introduction

In recent years, immunotherapy represented by immune checkpoint inhibitors (ICIs) has become a first-line treatment for multiple malignant tumors [1]. However, while exerting anti-tumor effects, ICIs may disrupt the normal immune tolerance balance in various organ systems including skin, colon, and lung, leading to immune-related adverse events (irAEs). The incidence of irAEs can be as high as 96%, with severe irAEs occurring in up to 55% of patients [2]. irAEs can occur throughout the entire treatment process and, when severe, can cause significant declines in organ function and quality of life, even becoming life-threatening [3]. Patient-reported outcome (PRO) refers to patients' subjective perceptions of their physical, psychological, social, and functional well-being, as well as overall health status, not constrained by objective differences such as laboratory reports or subjective judgments of healthcare providers, serving as an important indicator for measuring patients' experiences with irAEs and overall health status [4]. Therefore, evaluating PRO indicators for cancer patients undergoing immunotherapy is crucial in clinical efficacy assessments and drug trial reports [5], as it enables precise evaluation of irAEs following immunotherapy, facilitates early identification of irAEs in cancer patients, and supports timely development and implementation of targeted interventions, ensuring continued treatment and favorable prognosis [6].

Currently, patient-reported outcome measures (PROMs) commonly used for cancer immunotherapy are primarily generic scales with poor content validity, failing to measure nearly 30% of common irAEs [7] and unable to distinguish quality of life among cancer immunotherapy patients with different survival benefits [8]. At present, research on specific PROMs for cancer immunotherapy is in its infancy, with varying items and no unified standards, and lacks systematic evaluation of measurement properties. The Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) guideline is a tool that helps researchers select high-quality measurement instruments by evaluating the methodological quality and measurement properties of PROMs [9]. This study evaluates PROMs for cancer immunotherapy based on the COSMIN guideline to identify high-quality instruments and provide evidence-based support for accurately and effectively assessing irAEs and quality of life in cancer patients.

1. Materials and Methods

This study followed the 2024 version of the reporting guideline for systematic reviews of outcome measurement instruments (PRISMA-COSMIN) to report the measurement properties of the tools [10]. The study was registered in the PROSPERO database (registration number: CRD42023458328).

1.1 Search Strategy Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) principles, we conducted searches using a combination of subject headings and free-text terms in CNKI, Wanfang Data, SinoMed, PubMed, Embase, CINAHL, and ProQuest databases from inception to December 31, 2024. We also manually searched the reference lists of included studies. English search terms included "immunotherapy," "patient reported outcome measures," "index," and "development." Chinese search terms included "免疫治疗" (immunotherapy), "患者报告结局" (patient-reported outcome), and "量表" (scale). The search strategy for PubMed is shown in Table 1 , which was adapted from the methodological PubMed search filter developed by Terwee et al. [11] for finding studies on measurement properties of measurement instruments.

1.2 Inclusion and Exclusion Criteria **Inclusion criteria:** (1) Study subjects were adult cancer patients receiving ICIs; (2) Study content involved evaluation of at least one measurement property of PROMs for tumor immunotherapy; (3) Chinese or English language publications.

Exclusion criteria: (1) Studies where PROMs were used only as outcome measures or as validity criteria for other PROMs; (2) Reviews, commentaries, case reports, conference abstracts, and other non-primary research formats; (3) Studies where full text was unavailable or duplicate publications.

1.3 Literature Screening and Data Extraction Literature screening and data extraction were performed independently by two researchers who had received systematic training in evidence-based practice and COSMIN methodology, using a pre-designed data extraction form with cross-checking to ensure accuracy and completeness. Any disagreements were resolved through discussion or by a third researcher when necessary. Data extracted from the final included studies comprised: first author, publication year, country, sample size; PROMs name, study population, recall period, dimensions/number of items, language version; and measurement properties evaluated (content validity, structural validity, internal consistency, etc.).

1.4 Evaluation Methods

1.4.1 Evaluation Procedure According to the COSMIN guideline [12], literature quality evaluation was performed independently by two trained researchers, including methodological quality assessment and measurement property evaluation, with cross-checking and disagreement resolution through

discussion or third-party adjudication when necessary. First, the methodological quality of included studies was evaluated using the Chinese version of the COSMIN Risk of Bias checklist [13]. Then, the Chinese version of the COSMIN content validity scoring system [14] and the updated criteria for good measurement properties provided by COSMIN [12] were used to evaluate the measurement properties of PROMs, ensuring accuracy and consistency. Finally, evidence was synthesized and the quality of evidence was evaluated using the modified quantitative system for evidence grading [15] to form recommendation levels and recommendations for the measurement tools.

1.4.2 Evaluation Tools The evaluation comprised three components: methodological quality assessment, measurement property evaluation, and evidence grading assessment.

(1) Methodological quality assessment: The Chinese version of the COSMIN Risk of Bias checklist [13] was used to evaluate the methodological quality of included studies, covering 10 modules: PROMs development, content validity, structural validity, internal consistency, cross-cultural validity, reliability, measurement error, criterion validity, hypothesis testing, and responsiveness. Each measurement property module contains 3-35 items, with each item rated as "very good (V)," "adequate (A)," "doubtful (D)," "inadequate (I)," or "not applicable." According to the lowest score principle, the lowest rating among items represents the methodological quality rating for that measurement property. Items rated as "not applicable" were excluded from this principle.

(2) Measurement property evaluation: The Chinese version of the COSMIN content validity scoring system [14] was used to evaluate the measurement quality of PROMs content validity, including three aspects: relevance, comprehensiveness, and comprehensibility. The Chinese version of the updated criteria for good measurement properties provided by COSMIN [12] was used to evaluate other measurement properties of PROMs, including structural validity, internal consistency, reliability, measurement error, hypothesis testing, cross-cultural validity, criterion validity, and responsiveness. All measurement property ratings included "adequate (+)," "inadequate (-)," and "uncertain (?)." If a measurement property received inexplicable inconsistent ratings across studies, the overall rating for that property was "inconsistent ($\pm$)."

(3) Evidence grading assessment: The modified quantitative system for evidence grading [15] was used to evaluate the quality of evidence, including four factors: risk of bias, inconsistency, imprecision, and indirectness. Recommendation levels included "high," "moderate," "low," and "very low." The evaluation started by assuming "high" quality of evidence for each measurement property of PROMs, then downgraded based on the four factors above. Final recommendation levels were formed based on the overall rating of measurement properties and quality of evidence. Grade A indicates recommended use, where PROMs have "adequate (+)" content validity and at least low-level evidence supporting "adequate (+)" internal consistency. Grade B indicates potential for application,

suggesting further research is needed to validate quality, for PROMs that do not meet Grade A or C criteria. Grade C indicates not recommended, where high-level evidence proves any measurement property of the scale is "inadequate (-)."

---

2. Results

2.1 Literature Screening Results The initial search retrieved 2,280 articles: CNKI (n=31), Wanfang (n=0), SinoMed (n=3), PubMed (n=1,512), Embase (n=688), CINAHL (n=2), and ProQuest (n=44). Three additional articles were identified from reference lists. After removing duplicates, 1,569 articles remained. Following application of inclusion and exclusion criteria, 9 articles [16-24] were ultimately included. The literature screening flowchart is shown in Figure 1 [Figure 1: see original paper].

2.2 Basic Characteristics of Included Literature and Tools Among the 9 included studies, 7 [16,18-20,22-24] involved development and psychometric testing of immunotherapy PROMs, while 2 [17,21] were cross-cultural adaptation and validation studies. Considering that different language versions of scales differ in item content, adapted scales were treated as separate assessment tools. Therefore, a total of 9 immunotherapy PROMs were evaluated: Functional Assessment of Cancer Therapy-Immune Checkpoint Modulator (FACT-ICM), Chinese version of FACT-ICM (C-FACT-ICM), Patient-Reported Outcome for Fighting Financial Toxicity (PROFFIT), Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE) lung cancer subscale, MD Anderson Symptom Inventory Module Specific to Immunotherapy for Early-phase Trials (MDASI-Immunotherapy EPT), Chinese version of MDASI-Immunotherapy EPT, lung cancer immunotherapy PRO-CTCAE subset, Cancer Immunotherapy Patient Symptom Self-Report Scale, and Self-Report Symptom Inventory of Immune-Related Adverse Events in Patients with Lung Cancer (SRSI-irAEs-LC). The development period for these scales ranged from 2020-2024, with item counts ranging from 8-116. Two studies [18-19] included cancer patients receiving various treatments including chemotherapy and immunotherapy, while the remaining 7 studies [16-17,20-24] focused specifically on cancer immunotherapy patients. The basic characteristics of included literature and PROMs are shown in Table 2 .

2.3 Evaluation of Measurement Properties and Methodological Quality of Included Tools None of the 9 included studies reported cross-cultural validity, measurement error, or responsiveness. The methodological and measurement property quality evaluations for other properties are detailed below.

2.3.1 Validity Evaluation: Including Content Validity, Structural Validity, and Criterion Validity **2.3.1.1 Content Validity:** Three studies

[16,18-19] consulted experts and target patients regarding relevance, comprehensiveness, and comprehensibility, and modified the measurement tools based on evaluation results, receiving "very good" methodological quality ratings. Among these, two tools [16,18] were rated as having "adequate" content validity, while PRO-CTCAE-LC [19] was rated as "uncertain" due to failure to mention the recall period. Two studies [17,21] had insufficient comprehensiveness scoring: one [17] conducted cognitive interviews with 10 patients regarding content relevance, comprehensiveness, and comprehensibility, but only consulted 6 oncology experts about language appropriateness and content relevance/comprehensiveness; the other [21] only surveyed 8 experts and 20 immunotherapy cancer patients using quantitative methods. Therefore, these two studies [17,21] received "doubtful" methodological quality ratings for content validity, and their scales were rated as "uncertain." Four studies [20,22-24] did not clearly describe their content validity testing process in the text, receiving "inadequate" methodological quality ratings and "uncertain" content validity ratings for their tools.

2.3.1.2 Structural Validity: Five studies [17-18,21,23-24] evaluated structural validity. Three studies [17,23-24] used classical test theory and confirmatory factor analysis (CFA) to test scale structure, with adequate sample sizes, receiving "very good" methodological quality ratings. Two studies [17,24] reported root-mean-square error of approximation (RMSEA) <0.06, resulting in "adequate" structural validity ratings for their tools. Two studies [18,21] used exploratory factor analysis (EFA) but did not report fit indices, receiving "adequate" methodological quality ratings and "uncertain" structural validity ratings.

2.3.1.3 Criterion Validity: One study [24] reported criterion validity by calculating correlations using commonly used assessment tools as criteria, receiving an "inadequate" methodological quality rating but "adequate" criterion validity rating for the PROM due to correlation coefficients >0.7.

2.3.2 Reliability Evaluation: Including Internal Consistency and Test-Retest Reliability 2.3.2.1 Internal Consistency: Six studies [17-18,20-21,23-24] evaluated internal consistency (three studies [16,19,22] did not), all calculating internal consistency for each unidimensional scale or subscale, receiving "very good" methodological quality ratings. These six studies [17-18,20-21,23-24] provided sufficient evidence that Cronbach' s $\alpha \geq 0.7$ for each unidimensional scale or subscale, resulting in "adequate" internal consistency ratings.

2.3.2.2 Test-Retest Reliability: Three studies [17-18,24] reported test-retest reliability. Two studies [17-18] calculated intraclass correlation coefficient (ICC) values $ $0.7, but with time intervals of 3 weeks/21 days, exceeding the 2-week requirement in COSMIN guidelines, resulting in "inadequate" methodological quality ratings but "adequate" test-retest reliability ratings. One study [24] calculated Pearson correlation coefficient without evidence supporting absence of systematic change, but with a 1-week retest interval, receiving "inadequate" methodological quality rating and "uncertain" test-retest reliability rating.

2.3.3 Hypothesis Testing Three studies [17,21,24] evaluated hypothesis testing for structural validity using convergent and discriminant validity in similar populations, with >75% of results supporting hypotheses, receiving "very good" methodological quality ratings and "adequate" hypothesis testing ratings.

2.4 Quality of Evidence and Recommendations for Included Tools Regarding risk of bias, four studies [20,22-24] had "inadequate" methodological quality for content validity, resulting in a 2-level downgrade for content validity evidence quality; two studies [17,21] had "doubtful" methodological quality, resulting in a 1-level downgrade; the remaining three studies [16,18-19] had "very good" or "adequate" quality and were not downgraded. Five studies [17-18,21,23-24] had "very good" or "adequate" structural validity and were not downgraded. One study [24] had "inadequate" methodological quality for criterion validity, resulting in a 2-level downgrade. Six studies [17-18,20-21,23-24] had "very good" or "adequate" internal consistency and were not downgraded. Three studies [17-18,24] had "inadequate" methodological quality for test-retest reliability, resulting in a 2-level downgrade. Three studies [17,21,24] had "very good" methodological quality for hypothesis testing and were not downgraded. Regarding imprecision, three studies [16,19,22] only involved content validity, while other studies had sample sizes >100 and were not downgraded. Regarding indirectness, two studies [18,19] included patients receiving treatments other than immunotherapy, resulting in a 1-level downgrade for these measurement properties, while others were not downgraded. No inconsistent results were reported for any measurement tool, so no downgrades were applied for inconsistency.

In summary, according to COSMIN guidelines, content validity was "adequate" for FACT-ICM and PROFFIT, and "uncertain"for other scales. For internal consistency, FACT-ICM was not validated, and PROFFIT was rated "inadequate." None of the 9 tools [16-24] met the criteria for Grade A recommendation (adequate content validity plus high-level evidence of adequate internal consistency). Except for PROFFIT' s internal consistency having moderate-level evidence of being "inadequate," no tools had any measurement property rated as "inadequate" with high-level evidence, so no Grade C recommendations were assigned. All nine tools received Grade B recommendations. The quality of evidence and recommendations for included tools are shown in Table 4 .

---

3. Discussion

This study systematically reviewed 9 articles on cancer immunotherapy PROMs and evaluated their methodological quality and measurement properties according to COSMIN requirements. The results revealed existing problems in current research and related PROMs, with insufficiently detailed reporting and incomplete consideration of reliability and validity. This study summarizes these systematic review issues to provide references for future development or adaptation

of cancer immunotherapy PROMs.

3.1 Methodological Quality of Cancer Immunotherapy PROMs Needs Improvement The included studies had incomplete consideration of content validity, with qualitative methods needing improvement. Content validity is the most important measurement property of assessment tools [12]. According to COSMIN guidelines, content validity must be evaluated from both patient and expert perspectives regarding relevance, comprehensiveness, and comprehensibility. However, some included scales [17,19-20,23-24] did not comprehensively consider content validity during development, mostly relying on expert consultation or panel reviews with limited involvement of patients' subjective perspectives, affecting overall scale quality. Li et al. [25] and Zhou et al. [26] reported similar findings in their evaluations of quality assessment tools for cancer palliative care and fatigue assessment tools for cancer patients, noting that most scales primarily used expert consultation for content validity evaluation with minimal patient involvement.

Additionally, two included studies [21-22] used only quantitative surveys without qualitative research when developing scales, possibly due to time and resource constraints. Some studies may have focused more on quantitative data collection and analysis, neglecting the importance of qualitative interviews. Zhang et al. [27] similarly found that scales lacked qualitative methods in their evaluation of fear of cancer recurrence scales. Qualitative interviews provide indispensable in-depth information, helping researchers better understand patient experiences and needs. For example, Yang et al. [28] conducted three rounds of cognitive interviews with adult cancer patients when developing PRO measurement information system anxiety and depression scales for cancer patients, effectively eliminating scale gaps and completing professional and comprehensible semantic transformations to resolve target population understanding biases. Future development of cancer immunotherapy PROMs should include cognitive interviews with more than 7 patients by at least 2 experienced researchers to understand target population perspectives and ensure harmony between scale items and measured content or behaviors. Additionally, reporting should strictly follow COSMIN guidelines and qualitative research reporting standards [29] to improve comprehensiveness and standardization of patient-reported outcomes.

Criterion validity distinction was unclear in included studies, easily confused with hypothesis testing. Criterion validity assesses the extent to which PROMs results reflect a "gold standard" [26], but can be confused with hypothesis testing. The main difference lies in the comparison tools used: hypothesis testing examines correlations with other commonly used assessment tools or differences between subgroups [30]. However, researchers failed to clearly distinguish these methods in reporting, or did not adequately consider their differences during study design. For example, Fan et al. [24] selected commonly used assessment tools for correlation testing and provided hypotheses when developing SRSI-irAEs-LC, but reported it as criterion validity, which does not comply with

COSMIN guidelines. Future development of immunotherapy PROMs should review relevant literature, particularly COSMIN guidelines, to ensure study design and reporting meet guideline requirements, with detailed descriptions of methods and results for criterion validity and hypothesis testing, clearly identifying comparison tools to avoid confusion. Additionally, for abbreviated scales, COSMIN guidelines recommend using the original scale as a "gold standard" to evaluate criterion validity.

Test-retest reliability reporting in included studies was unclear, with retest methods needing optimization. Test-retest reliability is a fundamental tool for measuring indicator reliability and consistency [31]. COSMIN guidelines state that test-retest reliability should be conducted under the same test conditions with the same subjects, calculating ICC or Kappa values. However, not all included results reported test-retest reliability, or the reported design was unclear, possibly because researchers focused more on other psychometric properties. Additionally, three studies [17-18,24] violated COSMIN' s 2-week interval requirement: two studies [17-18] used 3-week intervals, which are too long to guarantee stability, while one [24] used a 1-week interval, which may cause memory interference and distorted reliability results. Similar findings were reported by Zhang et al. [32] and Lu et al. [33] in their evaluations of resilience scales for cancer patients and self-report outcome scales for liver cancer patients. Research shows that changes in measurement context or overly long/short intervals reduce reliability [34]. Therefore, future psychometric testing of cancer immunotherapy PROMs should improve test-retest reliability study design, set appropriate intervals, and conduct retests in populations and contexts similar to the initial measurement to ensure stable and reliable scale items.

3.2 Measurement Property Reporting for Cancer Immunotherapy PROMs Is Incomplete, Requiring Further Validation Although two assessment tools were cross-culturally adapted and validated from foreign work, they did not evaluate or report cross-cultural validity, which may affect reliability across cultural contexts. Future introduction of foreign cancer immunotherapy PROMs should evaluate cross-cultural validity by examining differential item functioning to ensure applicability and scientific validity.

Furthermore, none of the 9 included studies reported measurement error or responsiveness. Measurement error includes systematic and random error, representing changes other than true changes in the construct being measured [35]. According to COSMIN guidelines, measurement error for quantitative data should be assessed by calculating standard error of measurement through retesting, or by calculating limits of agreement and minimal detectable change; for binary/multicategorical/ordinal data, percentage agreement is recommended [36]. Responsiveness refers to a tool' s ability to detect changes in the measured construct over time, reflecting sensitivity to detect small changes and determining whether the tool can identify differences between subjects or within the same subject across stages [35]. Currently, development and psychometric testing of

cancer immunotherapy PROMs are still in early stages, and future research should include evaluation of measurement error and responsiveness to enhance scientific rigor.

For other measurement properties, three included studies [16,19,22] only reported content validity without addressing structural validity or internal consistency. Structural validity examines the degree to which scale structure aligns with the target construct and is an important indicator for evaluating overall scale structure [37], closely related to internal consistency evaluation results [14]. Content validity and internal consistency together determine measurement tool quality ratings [15]. Therefore, tool development should simultaneously consider internal consistency and structural validity, using exploratory or confirmatory factor analysis for comprehensive evaluation. Gao et al. [38] used both EFA and CFA in developing an exercise rehabilitation adherence scale for chronic heart failure patients, retaining highly sensitive and representative items. Future development of cancer immunotherapy PROMs can reference this approach, using EFA to infer factor structure during initial development and CFA after adjustment to ensure validity and reliability. Other measurement properties should also be evaluated to ensure scientific and rigorous item selection.

3.3 C-FACT-ICM Can Be Temporarily Recommended, But Requires Further Validation of Psychometric Properties In summary, this study comprehensively evaluated the psychometric properties of existing cancer immunotherapy PROMs and methodological quality of related studies based on COSMIN guidelines. The results show that C-FACT-ICM can be temporarily recommended (Grade B recommendation). The 9 included tools assessed cancer immunotherapy PRO from multiple dimensions, covering rich measurement perspectives. The lung cancer immunotherapy PRO-CTCAE subset [22] had the most items and broadest coverage, but some items were redundant, making completion cumbersome and time-consuming, and only content validity was verified with "inadequate" methodological quality, limiting precision of evaluation results. PRO-CTCAE-LC [19] had the fewest items and was simple to operate, but included all cancer patients, potentially limiting applicability to immunotherapy patients.

Based on evaluation of psychometric properties, methodological quality, and evidence grading, all included scales received Grade B recommendations. For content validity, although FACT-ICM [16] and PROFFIT [18] were rated "adequate," FACT-ICM [16] only validated content validity, while PROFFIT [18] had "inadequate" internal consistency, requiring further investigation of result stability and reliability. Other scales had "uncertain" content validity, indicating need for improvement in content validity and internal consistency research. For structural validity, only C-FACT-ICM [17] and SRSI-irAEs-LC [24] were rated "adequate," with others rated "doubtful" or "inadequate." C-FACT-ICM [17] received relatively comprehensive psychometric evaluation with moderate- to high-quality evidence supporting content validity, structural validity, and inter-

nal consistency. Developed by Meng et al. [17] in 2023 based on the FACT-ICM [16] from Princess Margaret Cancer Centre, University of Toronto, C-FACT-ICM includes two subscales (FACT-G and ICM) covering physical, emotional, social/family, functional well-being, and immune checkpoint modulator-specific modules, with 42 items total using a 5-point Likert scale (0-4), where higher scores indicate better quality of life. It effectively evaluates various experiences of cancer immunotherapy patients, though it has not yet been clinically applied. Overall, C-FACT-ICM demonstrates good psychometric properties for evaluating physical, psychological, and social functioning and quality of life in cancer immunotherapy patients and can be temporarily recommended, but it has not reported cross-cultural validity, measurement error, or responsiveness, requiring further supplementation. Its methodological quality and content validity measurement properties also need improvement, and the tool requires further clinical application to verify its usability and broad applicability.

This study has several limitations: (1) Only Chinese and English literature were included; (2) Studies not using measurement properties specified in COSMIN guidelines were excluded; (3) Newly developed scales without reported measurement properties were not included, potentially introducing bias. Future research should include more studies evaluating measurement properties of cancer immunotherapy PROMs, use other assessment tools for scale evaluation, and follow up on articles not reporting scale measurement properties or contact developers for information to improve comprehensiveness and representativeness.

References

[1] BAGCHI S, YUAN R, ENGLEMAN E G. Immune checkpoint inhibitors for the treatment of cancer: clinical impact and mechanisms of response and resistance[J]. Annu Rev Pathol, 2021, 16: 223-249. DOI:10.1146/annurev-pathol-042020-042741.

[2] CHAN K K, BASS A R. Autoimmune complications of immunotherapy: pathophysiology and management[J]. BMJ, 2020, 369: m736. DOI:10.1136/bmj.m736.

[3] OWEN C N, BAI X, QUAH T, et al. Delayed immune-related adverse events with anti-PD-1-based immunotherapy in melanoma[J]. Ann Oncol, 2021, 32(7): 917-925. DOI:10.1016/j.annonc.2021.03.204.

[4] GADGEEL S M. Patient-reported outcomes in the era of immunotherapy trials[J]. J Thorac Oncol, 2021, 16(4): 516-518. DOI:10.1016/j.jtho.2021.02.014.

[5] LIU M R, YAO M, ZHOU H, et al. Interpretation of the "Industry Guidelines for the Application of Core Patient-Reported Outcomes in Oncology Clinical Trials (Draft)" [J]. Chinese Journal of New Drugs, 2023, 32(7): 719-723. DOI:10.3969/j.issn.1003-3734.2023.07.010.

[6] BRAHMER J R, LACCHETTI C, SCHNEIDER B J, et al. Management of immune-related adverse events in patients treated with immune checkpoint inhibitor therapy: American society of clinical oncology clinical practice guideline[J]. J Clin Oncol, 2018, 36(17): 1714-1768. DOI:10.1200/JCO.2017.77.6385.

[7] COLOMER-LAHIGUERA S, BRYANT-LUKOSIUS D, RIETKOETTER S, et al. Patient-reported outcome instruments used in immune-checkpoint inhibitor clinical trials in oncology: a systematic review[J]. J Patient Rep Outcomes, 2020, 4(1): 58. DOI:10.1186/s41687-020-00210-z.

[8] VOON P J, CELLA D, HANSEN A R. Health-related quality-of-life assessment of patients with solid tumors on immuno-oncology therapies[J]. Cancer, 2021, 127(9): 1360-1368. DOI:10.1002/cncr.33457.

[9] GORST S L, PRINSEN C A C, SALCHER-KONRAD M, et al. Methods used in the selection of instruments for outcomes included in core outcome sets have improved since the publication of the COSMIN/COMET guideline[J]. J Clin Epidemiol, 2020, 125: 64-75. DOI:10.1016/j.jclinepi.2020.05.021.

[10] ELSMAN E B M, MOKKINK L B, TERWEE C B, et al. Guideline for reporting systematic reviews of outcome measurement instruments (OMIs): PRISMA-COSMIN for OMIs 2024[J]. J Clin Epidemiol, 2024, 173: 111422. DOI:10.1016/j.jclinepi.2024.111422.

[11] TERWEE C B, JANSMA E P, RIPHAGEN I I, et al. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments[J]. Qual Life Res, 2009, 18(8): 1115-1123. DOI:10.1007/s11136-009-9527-0.

[12] PRINSEN C A C, MOKKINK L B, BOUTER L M, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures[J]. Qual Life Res, 2018, 27(5): 1147-1157. DOI:10.1007/s11136-018-1798-3.

[13] MOKKINK L B, DE VET H C W, PRINSEN C A C, et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures[J]. Qual Life Res, 2018, 27(5): 1171-1179. DOI:10.1007/s11136-017-1765-4.

[14] SHEN L J, PENG J, CHEN Y T, et al. Introduction to COSMIN method: a scoring system for evaluating content validity of patient-reported outcome measurement tools[J]. Evidence-Based Nursing, 2021, 7(5): 609-614. DOI:10.12102/j.issn.2095-8668.2021.05.007.

[15] CHEN Y T, SHEN L J, PENG J, et al. Evaluation of patient-reported outcome measurement tools using the modified quantitative system for evidence grading[J]. Nursing Journal of Chinese People' s Liberation Army, 2020, 37(10): 57-60. DOI:10.3969/j.issn.1008-9993.2020.10.014.

[16] HANSEN A R, ALA-LEPPILAMPI K, MCKILLOP C, et al. Development of the functional assessment of cancer therapy-immune checkpoint modulator (FACT-ICM): a toxicity subscale to measure quality of life in patients

with cancer who are treated with ICMs[J]. Cancer, 2020, 126(7): 1550-1558. DOI:10.1002/cncr.32692.

[17] MENG X M, SHANG M M, WANG Q, et al. Reliability and validity of the simplified Chinese version of the functional assessment of cancer therapy-immune checkpoint modulator[J]. Qual Life Res, 2023, 32(6): 1581-1593. DOI:10.1007/s11136-022-03209-2.

[18] RIVA S, ARENARE L, DI MAIO M, et al. Cross-sectional study to develop and describe psychometric characteristics of a patient-reported instrument (PROFFIT) for measuring financial toxicity of cancer within a public healthcare system[J]. BMJ Open, 2021, 11(10): e049128. DOI:10.1136/bmjopen-2021-049128.

[19] VELDHUIJZEN E, WALRAVEN I, BELDERBOS J. Selecting a subset based on the patient-reported outcomes version of the common terminology criteria for adverse events for patient-reported symptom monitoring in lung cancer treatment: mixed methods study[J]. JMIR Cancer, 2021, 7(3): e26574. DOI:10.2196/26574.

[20] MENDOZA T, SHESHADRI A, ALTAN M, et al. Evaluating the psychometric properties of the immunotherapy module of the MD Anderson symptom inventory[J]. J Immunother Cancer, 2020, 8(2): e000931. DOI:10.1136/jitc-2020-000931.

[21] WU X D, XIE J Y, LIN X M, et al. Translation and validation of Chinese version of MDASI immunotherapy for early-phase trials module: a cross-sectional study[J]. BMC Nurs, 2023, 22(1): 176. DOI:10.1186/s12912-023-01217-9.

[22] PENG N N, ZHANG X J, CHEN F Z, et al. Construction of the lung cancer immunotherapy PRO-CTCAE subset[J]. Journal of Nurses Training, 2023, 38(19): 1729-1734. DOI:10.16821/j.cnki.hsjx.2023.19.001.

[23] YAN H Y. Construction of a self-report symptom scale for cancer immunotherapy patients[D]. Suzhou: Soochow University, 2023. DOI:10.27351/d.cnki.gszhu.2023.002108.

[24] FAN T T, ZHU S Y, WANG H, et al. Development and validation of the self-report symptom inventory of immune-related adverse events in patients with lung cancer[J]. Asia Pac J Oncol Nurs, 2024, 11(12): 100603. DOI:10.1016/j.apjon.2024.100603.

[25] LI J L, LIU N, GUO J. Systematic evaluation of methodological quality and measurement properties of quality assessment tools for cancer palliative care based on COSMIN[J]. Chinese General Practice, 2024: 1-9. DOI:10.12114/j.issn.1007-9572.2023.0227.

[26] ZHOU H M, HE L, XU H, et al. Systematic evaluation of fatigue assessment tools for cancer patients based on COSMIN guidelines[J]. Chinese General Practice, 2024: 1-10. DOI:10.12114/j.issn.1007-957.2024.0523.

[27] ZHANG L L, CHEN H, LUO H, et al. Systematic evaluation of fear of cancer recurrence assessment tools based on consensus standards for health measurement tools[J]. Chinese General Practice, 2023, 26(17): 2138-2146. DOI:10.12114/j.issn.1007-9572.2022.0810.

[28] YANG C, HUANG Y S, WU F L, et al. Development of the Chinese version of the patient-reported outcomes measurement information system anxiety and depression scales for cancer patients based on cognitive interviews[J]. Journal of Nurses Training, 2021, 36(22): 2069-2072. DOI:10.16821/j.cnki.hsjx.2021.30.003.

[29] DOSSETT L A, KAJI A H, COCHRAN A. SRQR and COREQ reporting guidelines for qualitative studies[J]. JAMA Surg, 2021, 156(9): 875-876. DOI:10.1001/jamasurg.2021.0525.

[30] STUCKY B D, PEREIRA C C A, DE VET H C W, et al. Measurement in medicine: a practical guide[J]. Qual Life Res, 2012, 21(2): 371-373. DOI:10.1007/s11136-012-0123-9.

[31] LU J Y, WEI Z P, ZHOU W J, et al. Reliability study of literature evidence retrieval: from an evidence-based perspective[J]. Library and Information, 2021(6): 60-68. DOI:10.11968/tsyqb.1003-6938.2021092.

[32] ZHANG Y S, ZHANG J, XU C, et al. Systematic evaluation of resilience assessment tools for cancer patients based on COSMIN guidelines[J]. Chinese General Practice, 2024, 27(29): 3664-3671. DOI:10.12114/j.issn.1007-9572.2023.0717.

[33] LU S Y, LIU X, JIANG X X, et al. Systematic evaluation of measurement properties of disease-specific self-reported outcome assessment tools for liver cancer patients[J]. Chinese Journal of Nursing, 2024, 59(22): 2734-2741. DOI:10.3761/j.issn.0254-1769.2024.22.007.

[34] SAHANDI FAR M, EICKHOFF S B, GONI M, et al. Exploring test-retest reliability and longitudinal stability of digital biomarkers for parkinson disease in the m-power data set: cohort study[J]. J Med Internet Res, 2021, 23(10): e26608. DOI:10.2196/26608.

[35] LV P P, CUI N X, HAN J, et al. Systematic evaluation of anxiety assessment tools for children with autism spectrum disorder[J]. Chinese Journal of Child Health Care, 2022, 30(12): 1358-1363. DOI:10.11852/zgetbjzz2022-0271.

[36] PENG J, SHEN L J, CHEN Y T, et al. Interpretation of the COSMIN-RoB checklist for risk of bias in studies on measurement instrument stability, measurement error, and criterion validity[J]. Chinese Journal of Evidence-Based Medicine, 2020, 20(11): 1340-1344. DOI:10.7507/1672-2531.202003164.

[37] CAO L P, ZOU C J, ZHU K Y, et al. Translation and psychometric testing of the intermittent self-catheterization questionnaire[J]. Chinese Nursing Research, 2025, 39(7): 1132-1137. DOI:10.12102/j.issn.1009-6493.2025.07.012.

[38] GAO M, SUN G Z, WANG Q Y, et al. Development and psychometric testing of an exercise rehabilitation adherence scale for patients with chronic heart failure[J]. Chinese General Practice, 2024, 27(25): 3150-3158. DOI:10.12114/j.issn.1007-9572.2022.0081.

Author Contributions: SU Zhenzhen was responsible for conceptualization, data curation, manuscript writing, and revision; SU Zhenzhen and WANG Yixuan were responsible for literature collection and organization; ZHANG Liyan was responsible for quality supervision and review; SU Zhenzhen, LIAN Xuemin, and LIU Dan were responsible for data extraction.

Conflict of Interest: The authors declare no conflicts of interest.

ORCID: SU Zhenzhen https://orcid.org/0009-0001-0654-4864

Received: 2025-04-10
Revised: 2025-07-10
Accepted for publication: 2025-07-15

Note: Figure translations are in progress. See original paper for figures.