International Journal of Nursing Studies 79 (2018) 58–69

Contents lists available at ScienceDirect

International Journal of Nursing Studies

journal homepage: www.elsevier.com/locate/ijns

Evaluating test-retest reliability in patient-reported outcome measures for older people: A systematic review

Myung Sook Park[a], Kyung Ja Kang[b], Sun Joo Jang[c], Joo Yun Lee[d], Sun Ju Chang[e,*]

[a] Nursing Department, Konkuk University, Chungju, South Korea
[b] College of Nursing, Jeju National University, Jeju, South Korea
[c] College of Nursing, Eulji University, Daejeon, South Korea
[d] The Research Institute of Nursing Science, Seoul National University, Seoul, South Korea
[e] College of Nursing & The Research Institute of Nursing Science, Seoul National University, Seoul, South Korea

ARTICLE INFO

Keywords:
Test-retest reliability
Patient-reported outcomes
Systematic review
Aged

ABSTRACT

Objectives: This study aimed to evaluate the components of test-retest reliability including time interval, sample size, and statistical methods used in patient-reported outcome measures in older people and to provide suggestions on the methodology for calculating test-retest reliability for patient-reported outcomes in older people.
Design: This was a systematic literature review.
Data sources: MEDLINE, Embase, CINAHL, and PsycINFO were searched from January 1, 2000 to August 10, 2017 by an information specialist.
Review methods: This systematic review was guided by both the Preferred Reporting Items for Systematic Reviews and Meta-Analyses checklist and the guideline for systematic review published by the National Evidence-based Healthcare Collaborating Agency in Korea. The methodological quality was assessed by the Consensus-based Standards for the selection of health Measurement Instruments checklist box B.
Results: Ninety-five out of 12,641 studies were selected for the analysis. The median time interval for test-retest reliability was 14 days, and the ratio of sample size for test-retest reliability to the number of items in each measure ranged from 1:1 to 1:4. The most frequently used statistical methods for continuous scores was intraclass correlation coefficients (ICCs). Among the 63 studies that used ICCs, 21 studies presented models for ICC calculations and 30 studies reported 95% confidence intervals of the ICCs. Additional analyses using 17 studies that reported a strong ICC (> 0.09) showed that the mean time interval was 12.88 days and the mean ratio of the number of items to sample size was 1:5.37.
Conclusions: When researchers plan to assess the test-retest reliability of patient-reported outcome measures for older people, they need to consider an adequate time interval of approximately 13 days and the sample size of about 5 times the number of items. Particularly, statistical methods should not only be selected based on the types of scores of the patient-reported outcome measures, but should also be described clearly in the studies that report the results of test-retest reliability.

What is already known about the topic?

- Current literature has proposed common factors and quality criteria for evaluating the psychometric properties of patient-reported outcome measures.
- Of the psychometric properties, the test-retest procedure used to assess stability is exposed to several risks, such as carryover effects and actual change between two separate times.
- Although the time interval for test-retest reliability for older people might be different from that of the general population, current literature related to the quality of test-retest reliability including time interval, sample size, and statistical methods used in patient-reported measures for older people has not been evaluated yet.

What this paper adds

- The median time interval between two administrations was 14 days.
- The mean time interval and the mean ratio of the number of items in each measure to sample size for test-retest reliability in studies that reported a strong Intraclass correlation coefficient (ICC) were

* Corresponding author at: College of Nursing Seoul National University, Daehak-ro 103, Jongro-gu, Seoul, 406-799, South Korea.
E-mail addresses: elderly1004@hanmail.net (M.S. Park), kkyungja@jejunu.ac.kr (K.J. Kang), icedcoffee@eulji.ac.kr (S.J. Jang), jylee3130@gmail.com (J.Y. Lee), changsj@snu.ac.kr (S.J. Chang).

https://doi.org/10.1016/j.ijnurstu.2017.11.003
Received 30 June 2017; Received in revised form 7 November 2017; Accepted 7 November 2017

12.88 days and 1:5.37, respectively.
- Most of studies that used continuous scores for test-retest reliability evaluated the reliability using the ICC; however, less than half these studies reported models for calculation and 95% confidence intervals of the ICC.

1. Introduction

With a great interest in concepts related to patient-centered care in the health care system, patient-reported outcomes have been emphasized around the world (Adler and Resnick, 2010). patient-reported outcomes, which indicate all kinds of information coming from patients, have been widely used for a variety purposes including evaluating health care quality, screening health risks and problems, and assessing the effects of treatment or interventions (Adler and Resnick, 2010; Deshpande et al., 2011; Nelson et al., 2015). Although the scientific knowledge about the impact of patient-reported outcomes is still debatable, the use of patient-reported outcomes in the health care system has rapidly emerged because healthcare providers can directly hear the patient's voice and obtain value from hearing the patient's perspective (Adler and Resnick, 2010; Nelson et al., 2015; Santana and Feeny, 2014).

Given the emphasis on patient-reported outcomes, attention needs be given to the measures that assess patient-reported outcomes because further plans and actions related to treatments or interventions could be changed depending on the findings of a patient-reported outcome measure (Frost et al., 2007). Hence, the psychometric properties of patient-reported outcome measures must be ensured (Deshpande et al., 2011; Frost et al., 2007; Nelson et al., 2015). Current literature has proposed common factors and quality criteria for evaluating the psychometric properties of patient-reported outcome measures; those factors are validity and reliability (Deshpande et al., 2011; Frost et al., 2007; Nelson et al., 2015; Terwee et al., 2007).

Validity is the extent to which a measure accurately measures what it is intended to measure (DeVellis, 2012; Streiner and Norman, 2008; Waltz et al., 2010). Three fundamental types of validity have been widely used to evaluate the validity of a patient-reported outcome measure: content validity refers to whether the items of a measure represent the content domain, construct validity refers to whether a measure correlates with theoretical concepts it is supposed to be related to as well as not be related to, and criterion-related validity refers to whether a measure correlates with a "gold standard" measure as a criterion (DeVellis, 2012; Streiner and Norman, 2008; Waltz et al., 2010). Reliability is the extent to which a measure is able to provide consistent and accurate results related to the target attribute (DeVellis, 2012; Polit and Beck, 2008; Waltz et al., 2010). For estimating reliability, three procedures are commonly used: internal consistency, which refers to the coherence of items within a measure; equivalence, which concerns the degree of agreement among two or more observers; and stability, which concerns obtaining comparable results at two separate times (DeVellis, 2012; Polit and Beck, 2008; Waltz et al., 2010).

Among the psychometric properties evaluating patient-reported outcome measures, the test-retest procedure used to assess stability is exposed to several risks, such as carryover effects and actual change between two separate times (DeVellis, 2012; Polit and Beck, 2008; Yu, 2005). These risks could be minimized by determining the appropriate interval between two administrations (Deshpande et al., 2011; Streiner and Norman, 2008). A short time interval might cause recall of the items, and a long time interval might permit clinical change over the time period (DeVellis, 2012; Terwee et al., 2007). Two to 14 days between the first and second administrations are generally acceptable for evaluating test-retest reliability (Streiner and Norman, 2008; Terwee et al., 2007; Waltz et al., 2010). However, the appropriate time interval could differ for diverse characteristics such as the target group's age (Frost et al., 2007; Streiner and Norman, 2008).

With rapidly emerging health issues related to aging, various patient-reported outcome measures have been developed and validated for older people. Given the changes in cognitive functioning such as memory, as well as health conditions in older people (Denton and Spencer, 2010; Gilsky, 2007), the appropriate time interval for evaluating test-retest reliability in older people might be different from that of the general population. Unfortunately, current literature related to the quality of test-retest reliability including time interval, sample size, and statistical methods used in patient-reported outcome measures in older people has not been evaluated yet. Therefore, this study aimed to evaluate the components of test-retest reliability including time interval, sample size, and statistical methods used in patient-reported outcome measures in older people, and to provide suggestions on the methodology for calculating test-retest reliability for patient-reported outcomes in older people.

2. Methods

This systematic review employed both the Preferred Reporting Items for Systematic Reviews and Meta-analysis (PRISMA) checklist (Equator Network, 2013) and the guideline for systematic review that was published by the National Evidence-based Healthcare Collaborating Agency (NECA) in Korea (Kim et al., 2011).

2.1. Search strategies

In order to identify the eligible studies, the electronic databases including MEDLINE, Embase, CINAHL, and PsycINFO were searched from January 1, 2000 to August 10, 2017 by an information specialist. The reason for the starting date of 2000 for the search was that the term patient-reported outcomes was suggested by the US Food and Drug Administration in 2001, although similar concepts were used from the 1970s to 1990s (Wu et al., 2013). The following search terms were determined in accordance with the PICO (population, index, comparison, and outcomes) model: population was "aged," "elderly," or "older adults"; index test was "test-retest reliability"; and comparisons and outcomes were not set for the research questions of this systematic review. The search strategies using the combined search terms in each database are delineated in Table 1. To find additional eligible studies, the researchers reviewed the reference lists of the selected studies.

Table 1
Search strategies.

	Ovid-MEDLINE	Ovid-EMBASE	CINAHL[a] complete	PsycINFO
1	Exp Aged/	Exp Aged/	(MH "Aged+")	Exp Aged/
2	elderly.mp.	elderly.mp.	(MH "Test-retest reliability")	elderly.mp.
3	"older adult$1".mp.	"older adult$1".mp.	1 AND 2	"older adult$1".mp.
4	OR/1–3	OR/1–3	limit 3 to yr = "2000–2017"	OR/1–3
5	"test-retest reliability".mp.	"test-retest reliability".mp.		"test-retest reliability".mp.
6	4 AND 5	4 AND 5		4 AND 5
7	limit 6 to yr = "2000–Current"	limit 7 to yr = "2000 − Current"		limit 7 to yr = "2000 − Current"
Results	3808	5097	3365	371

[a] CINAHL, Cumulative Index to Nursing and Allied Health Literature; The word "Current" in this table indicates between the first and second weeks in August 2017.