VALUE IN HEALTH ▮ (2018) ▮▮▮–▮▮▮

Available online at www.sciencedirect.com

ScienceDirect

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

Original Research

Evaluating the Content Validity of Four Performance Outcome Measures in Patients with Elective Hip Replacements and Hip Fractures

Rachel Ballinger, PhD[1,*], *Cicely Kerr, PhD*[2], *Fiona Mowbray, PhD*[2], *Elizabeth Nicole Bush, MHS*[3]

[1]*ICON Clinical Research, Abingdon, Oxfordshire, UK;* [2]*Formerly ICON Clinical Research, Oxford, Oxfordshire, UK;* [3]*Patient-Focused Outcomes Center of Expertise, Eli Lilly and Co., Indianapolis, IN, USA*

ABSTRACT

Objectives: To assess the content validity of performance outcome (PerfO) measures for use with patients undergoing hip fracture (HF) surgery and elective total hip replacement (eTHR). **Methods:** This study was a substudy of a broader evaluation of measurement properties of PerfO measures. The PerfO measures assessed were timed up and go (TUG), four-step stair climb (4SC), long stair climb (LSC), and repeated chair stand (RCS). For this substudy, HF and eTHR participants were interviewed to evaluate the relevance and difficulty of each PerfO measure. Qualitative analysis was conducted on interview transcripts, and summaries of coded data were produced to assess saturation. **Results:** All 18 HF participants related the PerfO measures (TUG, 4SC, and RSC) to activities they completed in daily life, with slight variations in some specific aspects. For the eight eTHR participants, the correspondence between the PerfO measures (TUG, 4SC, and LSC) and activities in daily life varied: all participants saw similarity in the movements for the TUG; most undertook short stair climbs in daily life, but most did not regularly undertake LSC in daily life. Nevertheless, all HF and eTHR participants reported that the PerfO measures were relevant and had a level of difficulty similar to daily activities. **Conclusions:** This study contributes novel methods that adapt US regulatory guidance for patient-reported outcome measures to the evaluation of PerfO measures. A structured approach was used to explore specific details of each measure and correspondence to everyday life. This study demonstrates how content validity of PerfO measures can be meaningfully assessed.

Keywords: content validation, hip fracture, hip replacement, performance outcomes.

Introduction

Performance outcome (PerfO) measures are a type of clinical outcome assessment (COA) outlined by the US Food and Drug Administration (FDA). A PerfO measure is based on tasks performed by a patient according to instructions; it is administered by a health care professional but requires patient cooperation and motivation [1]. PerfO measures can provide specific information about functional status and mitigate variance introduced by perception of functional ability [2]. Specifically in clinical trial evaluation of orthopedic treatment, PerfO measures have been used to assess functioning, including timed up and go, stair climb, chair stand, fast-paced walk, and 6-minute walk tests [3,4].

As with all types of COAs used in clinical trials of medical products, PerfO measures should reflect the health experiences of patients in terms of how they feel or function in everyday life [1,5]. Nevertheless, although some PerfO measures assess abilities and actions that closely simulate how a patient functions in typical life, others assess concepts of interest for which the connection to everyday activities is less clear, such as supine quadriceps isometric strength [6]. In the regulatory context, the degree of correspondence between the COA measurement concept and how patients feel or function in everyday life is considered a key element of content validity. The FDA has provided guidance for the assessment of content validity of patient-reported outcome (PRO) measures and defined *content validity* as the extent to which the PRO instrument measures the concept of interest [5]. Furthermore, qualitative evidence is required to demonstrate that the items and domains of an instrument are appropriate and comprehensive relative to its intended measurement concept, population, and use [5], and it must be based on direct input from an adequate sample of patients from the targeted clinical study population [7]. Such specific guidance in relation to content validation of PerfO

* *Address correspondence to:* Rachel Ballinger, ICON Clinical Research UK Ltd., 100 Park Drive, Milton Park, Abingdon, Oxon OX14 4RY, UK.

E-mail: Rachel.Ballinger@iconplc.com

1098-3015$36.00 – see front matter https://doi.org/10.1016/j.jval.2018.02.005

measures is not available. In addition, PRO measures are designed to directly capture patient experience, for example, how patients feel or function, and hence PRO measurement concepts link closely to meaningfulness to patients. For other types of COAs, the meaningfulness of the measurement concept to patients' everyday life may need to be considered differently or separately.

On the basis of literature review and expert clinical opinion, four PerfO measures were selected to assess performance in three study populations as part of the main evaluation study (reported elsewhere [8]). This substudy specifically assessed the content validity of four PerfO measures: timed up and go (TUG), four-step stair climb (4SC), long stair climb (LSC), and repeated chair stand (RCS). Assessment of content validity was based on FDA's 2009 industry guidance titled "Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims" (henceforth, PRO guidance) [5] to the extent that it could be applied to PerfO measures. In addition, specific feedback was received that indicated that the FDA was interested in knowing the relevance of the measures and how the measures' level of difficulty related to everyday functioning.

Methods

Study Design

This content validation study was a qualitative substudy of a main evaluation study, with a longitudinal design assessing the measurement properties of the same PerfO measures (Fig. 1). The main and substudy protocols were approved by a central institutional review board and all participants provided written informed consent before enrolling. The main PerfO measure evaluation study was conducted at 15 clinical sites in the United States and evaluated select PerfO measures in participants who underwent hip fracture (HF) surgery, elective total hip replacement (eTHR), or elective total knee replacement (eTKR). During each of the three main study visits, PerfO measures were administered by trained health care professionals and included TUG, 4SC, LSC, and RCS (described in Table 1). As a predictor of future falls [10], the RCS was undertaken only with the HF group. The LSC was used only with the eTHR and eTKR groups.

For the content validation substudy, HF and eTHR participants in the main study were invited to complete a telephone interview after one of their scheduled study visits; eTKR participants had completed all three visits at the time the substudy was initiated and so were not included. Interviews were conducted between November 2013 and May 2014. At that time, HF participant recruitment for the main study was ongoing, allowing inclusion of HF participants in the substudy who were at different stages postsurgery; nevertheless, the remaining eTHR participants were all attending their final main study visit 12 weeks (±3 days) postsurgery (Fig. 1).

Structured interview guides for each group were developed. These included a recap of the relevant PerfO measures and instructions given to help focus participants' recall. Questions related to overall experience and specific details of each measure (such as rising to standing and turning), before exploring the relevance of the measures to everyday activities and functioning.

Participants

Inclusion criteria for the content validation substudy were participation in the main evaluation study (eligibility criteria are presented in Table 2) and availability for a telephone interview, ideally within 5 days of a main study visit. With participant permission, selected main study data were made available to the interviewer for reference during the interview (e.g., if a test had not been fully completed).

Interviews

Telephone interviews were conducted by experienced interviewers using the relevant structured interview guide. Interviews were scheduled to last less than 45 minutes and were audio-recorded and transcribed verbatim. Participants were reimbursed for their time.

Analysis

Analysis codes were identified both from the interview guide and from themes that emerged directly from the data. A codebook for each patient group was developed after review of data from the first four interviews in each group; these were reviewed against transcripts by additional members of the study team. The codebooks comprised code names, definitions, and examples to help ensure consistency of coding across interviews. These documents were modified as needed during the coding of subsequent interviews (e.g., to reflect newly identified themes/codes). Qualitative analysis MAXQDA software (Sozialforschung GmbH, Berlin, Germany) was used to code data.

Sample size was limited by the availability of participants remaining on active follow-up in the main study at the time of the substudy data collection, particularly the eTHR group. Nevertheless, data saturation on core themes was thoroughly assessed in both groups to inform level of confidence in the results and conclusions.

Given the structured interview guides with focused exploration of correspondence between specific test movements and everyday life, the standard approach to assessing data saturation in concept elicitation studies [12] was not appropriate for this qualitative study. Instead, data summary grids were completed for each participant group and reviewed for saturation: interview content was summarized by participant and by PerfO measure for each of three core themes, drawing on the content of groups of analysis codes, namely, overall relevance, overall speed (relates to both relevance and difficulty because instructions varied between the PerfO measures assessed: normal walking speed or as fast as safely able), and overall level of difficulty. The content of new details identified from the final eTHR and the last three HF interviews was reviewed to assess the value of any new details identified at that point, to inform consideration of whether additional interviews would yield important additional information. This approach of summarizing new content rather than just indicating application of a new code is similar to that proposed by Brod et al. [13]. A team-based approach was used to develop and check the accuracy of the summary grid content details (as described in Fig. 1).

Results

Participant and Interview Characteristics

The study sample comprised 18 HF participants recruited from three sites (24% of the 75 HF participants at baseline in the main evaluation study) and 8 eTHR participants from five sites (9.5% of the 84 eTHR participants at visit 3 in the main evaluation study). All interviews were conducted within 7 days after the main study visit at which the participants had completed the PerfO measures (mean days after visit: HF, 3; eTHR, 4). HF participants were interviewed 79 to 177 days after surgery and across the three visits of the main evaluation study (following visit 1, n = 4; visit 2, n = 8; visit 3, n = 6). The mean length of the interviews was 36 minutes (range 18–50 minutes).

Fig. 1 – Evaluation of content validity substudy overview. *Main evaluation study also included participants with eTKR surgery. No eTKR participants took part in the content validation substudy because it took place after all eTKR main study visits had been completed. **Participants repeated the assessments within a visit to enable evaluation of inter-rater variation/reliability: eTHR at the first visits, for HF at the second (for timed up and go and four-step stair climb) and third (repeated chair stand) visits. eTHR, elective total hip replacement; eTKR, elective knee replacement; HF, hip fracture; PerfO, performance outcome.

Participant characteristics are presented in Table 3, including comparison with the main evaluation study sample. This shows that participants were broadly reflective of the main evaluation study sample. There were some moderate differences (10% or more) when comparing eTHR substudy participants with those in the main evaluation study: there were more participants who were female, employed full-time, with graduate degree (but fewer than in relation to other education levels), and who had comorbidities. HF participants who used a staircase at home were slightly over-represented in the substudy sample compared with the main evaluation study sample.

At the time of interview, 17 out of 18 HF participants felt they were doing well after their surgery. One HF participant felt he had not experienced much improvement after surgery; he was still in pain and required the use of an assistive device. All eight eTHR participants felt they were doing well 3 months after their hip surgery.

Participant Experience with Performing PerfO Measures' Activities/Movements

Example quotations from participants are presented in the first section of Table 4.

Timed up and go

All HF and eTHR participants found the TUG test feasible to complete overall, often reporting that this was due to lack of discomfort or pain (two HF and three eTHR participants), not

Table 1 – Descriptions of PerfO measures assessed.

PerfO measure description	Assessed in	
	HF	eTHR
Timed up and go (TUG): The TUG is a test of balance that is commonly used to examine physical performance and lower extremity strength in community-dwelling, frail, older adults [9]. The TUG test measures the time taken by an individual to stand up from a standard chair, walk a short distance at their normal walking speed, turn, walk back to the chair, and sit down. The participant is able to use any usual walking aid and should wear their usual shoes. The time (recorded in seconds) taken to complete the task is strongly correlated with the level of functional mobility, with faster time representing better performance.	✓	✓
Stair climb (SC)—4 steps: The SC test is a physical performance measure that assesses the time (seconds) it takes a patient to ascend a predetermined number of steps or as many as they are able to, as fast as they feel safe to do so. The test provides an indication of general functional ability and lower extremity muscle power. Participants were asked to climb up the four steps of a four-step portable staircase.	✓	✓
SC—12 steps: eTHR participants who completed the four-step SC were invited to undertake a separate test to ascend 12 steps up a usual staircase.	✗	✓
Repeated chair stand (RCS): The RCS test is a measure of leg strength and power and is recognized as a predictor of future falls [10,11]. In the 5-time RCS, patients are asked to rise from a chair 5 times, as fast as possible with their arms folded on their chest and to repeat the test using the chair armrests after a short rest in between. Performance was measured in seconds as the time from the initial seated position to the final standing position on arising from the chair for the fifth time.	✓	✗

eTHR, elective total hip replacement; HF, hip fracture; PerfO, performance outcome.

Table 2 – Inclusion and exclusion criteria for the main PerfO evaluation study in patients with HF or eTHR.

HF	eTHR
	Inclusion criteria
Males or females aged ≥65 y	Males or females aged ≥50 y
Unilateral proximal femur fracture, with noncomplicated surgical repair within 3–12 wk before baseline visit	Either primary eTHR because of hip OA or revision surgery (after failure of a primary eTHR that was originally performed because of hip OA) and is planned within 15 d to 8 wk after baseline visit
Ambulatory before the fracture (with or without assistive device)	
Able to stand up from a chair and walk more than five steps (~4 m) without human assistance (any assistive device allowed) according to the patient	
BMI <40 kg/m^2 or a weight <136.4 kg	
Informed consent by IRB-approved informed consent form	
	Exclusion criteria
HF resulting from a bone neoplasm or major trauma	Another inpatient lower limb surgical procedure planned in the 6 mo after baseline visit
Lower extremity amputation (foot, leg, or thigh)	Lower limb fracture within previous 6 mo
	Simultaneous bilateral eTHR
	The planned surgical procedure would preclude weight bearing for at least 4 wk postoperatively*
Underlying muscle disease (e.g., polymyositis or muscular dystrophy) or a history of muscle disease other than age-associated muscle waste or disuse atrophy	
Progressive disorder(s) likely to severely confound physical performance tests during the course of the study (such as unstable Parkinson disease, severe peripheral neuropathy, motor neurone disease, or hemiplegia)	
Severe psychiatric disorder or cognitive impairment that in the opinion of the investigator would interfere with protocol procedures	
Unable to safely perform the protocol-specified tests of physical performance because of comorbidity (e.g., visual/hearing impairment, MI, and pulmonary disease)	
Patients already participating in any trial whereby their mobility may be impacted	
Investigator site personnel directly affiliated with this study and/or their immediate families†	
Eli Lilly and Co. employees or employees of a designated third-party organization assisting with the conduct of the study	

BMI, body mass index; eTHR, elective total hip replacement; HF, hip fracture; IRB, institutional review board; MI, myocardial infarction; OA, osteoarthritis; PerfO, performance outcome.

* "Partial weight bearing" and "weight bearing as tolerated" were acceptable, but "non– weight-bearing," "touch weight bearing," or "feather weight bearing" were excluded.

† Immediate family was defined as a spouse, parent, child, or sibling, whether biological or legally adopted.

Table 3 – Demographic and medical history characteristics of participants of the qualitative substudy and the main evaluation study.

Demographic characteristic	Unit	HF participants (N = 18)	HF main evaluation study (N = 75)	eTHR participants (N = 8)	eTHR main evaluation study (N = 98)
Mean age (range)	Years	78.2 (65.6–86.7)	79.6 (65.5–94.6)	67 (57–78)	67.5 (50–91)
Sex, n (%)	Female/Male	12/6 (67/33)	51/24 (68/32)	6/2 (75/25)	64/34 (65/35)
Ethnicity, n (%)	White	17 (94)	71 (95)	8 (100)	92 (94)
	Other (%, specify)	1 (6, Asian)	4 (5; 3 Hispanic, 1 black, 1 other)	0	6 (6; 4 Hispanic, 1 black, 1 other)
Employment status, n (%)	Retired	15 (83)	64 (85)	5 (63)	54 (55)
	Employed part-time	1 (6)	4 (5)	0	9 (9)
	Unemployed (seeking work)	1 (6)	1 (1)	0	2 (2)
	Looking after home/family	1 (6)	4 (5)	0	1 (1)
	Full-time employment	0	2 (3)	3 (38)	22 (23)*
Education, n (%)	High school	6 (33)	27 (36)	3 (38)	28 (29)
	College degree	5 (28)	19 (25)	1 (13)	23 (24)
	Some college	4 (22)	13 (17)	1 (13)	35 (36)
	Graduate degree	2 (11)	9 (12)	2 (25)	9 (9)
	Did not complete high school	1 (6)	7 (9)	1 (13)	3 (3)
Use staircase at home, n (%)	No/Yes	11/7 (61/39)	57/18 (76/24)	5/3 (63/38)	59/39 (60/40)
Lower limb impediment, n (%)	None/Has impediment	12/6 (67/33)	52/23 (69/31)	4/4 (50/50)	52/44 (53/45)†
Comorbidities, n (%)	Have comorbidities	17 (94)	70 (93)	8 (100)	88 (90)

eTHR, elective total hip replacement; HF, hip fracture.

* Remaining 10 participants: 4 permanently unable to work, 3 temporarily unable to work, 2 other, and 1 a student.

† Two had other impediments (bilateral knee patellofemoral arthrosis and bone spurs).

needing any assistance device (three eTHR participants), finding it similar to daily activities (two HF participants), or being able to use chair armrests (two HF participants). Several eTHR and HF participants reported slight difficulty in some aspects of the TUG test, for example, standing up from the chair (three HF and one eTHR participant), turning (one HF and one eTHR participant), balance (one HF participant), or walking (one eTHR participant).

Four-step stair climb

Overall, most HF participants (n = 15) and all eTHR participants found the 4SC test feasible to complete at their most recent visit. Reasons given for the ease of the test included the test being an activity that they did regularly in daily life (two HF participants), no longer having any pain when climbing steps (two HF participants), the surface of the steps being stable and clear (one HF participant), climbing a few steps rather than a longer flight (two HF participants), increased confidence (one eTHR participant), and appropriate step height (three eTHR participants).

Repeated chair stand

All HF participants were able to complete the RCS test when using chair armrests, with 17 reporting that the test was feasible to complete. Reasons given for the ease included lack of pain or discomfort (two participants) and the ability to use the chair armrests for support and to push themselves to stand up (six participants). Twelve HF participants were also able to complete the RCS test with their arms folded, seven of whom said both versions of the RCS test were feasible to complete.

Long stair climb

Overall, the eight eTHR participants did not have any difficulty in completing the LSC test. Six eTHR participants reported that they had no difficulty with either lifting their feet up or pushing down with their feet when stepping up. One participant, however, found the motion of lifting up each leg and placing it on the next step slightly more difficult than the other aspects of the test; another found that her hip became stiffer by the time she reached the last steps, and another reported that his legs became tired toward the end of the climb.

Extent to Which the PerfO Measures' Activities/Movements Reflect Ability to Function in Daily Life

Example quotations from participants are presented in the second section of Table 4.

Timed up and go

All participants were able to relate the overall TUG test to movements that they engaged in regularly and all reported that the test was an accurate and relevant reflection of how they performed similar movements in their everyday lives. Only slight

Table 4 – Example quotes.

PerfO	eTHR	HF
	Patient experiences of completing the PerfO assessments	
Timed up and go	"Yes, it was easy because I didn't need any help aids or anything." (R1, female, 61 y old, 87 d since surgery)	"Standing up was all right once I got up. It was getting up at the time." (F5, male, 86 y old, 120 d since fracture)
Four-step stair climb	"I think just initially my thrust going up probably the first step or two, since I had not been doing that for a little bit over three months. Maybe that was a little bit harder, but not much." (R5, female, 72 y old, 89 d since surgery)	"I had no difficulty climbing the steps, but I did have to, you know, use the handrails." (F11, female, 79 y old, 177 d since fracture)
Long stair climb (eTHR)/ Repeated chair stand (HF)	"I might have went faster, but I slowed down right at the top because my legs were tired." (R2, male, 69 y old, 84 d since surgery)	"Well, you're using your legs more, and you're using your leg muscles to get up, and your hip, and everything as opposed to your arms. All your effort is on your legs, which is what it should be. If you're going to improve you should be using those muscles." (F10, female, 65 y old, 167 d since fracture)
	Patient views on the extent to which PerfO assessments reflect ability to function in daily life	
Timed up and go	"I tried to act as normal as possible. Like if I would get up from a chair at home and start walking, I would do it the same way." (R6, female, 78 y old, 90 d since surgery)	"It's something you're doing all the time, getting up and down, walking around things." (F10, female, 65 y old, 167 d since fracture)
Four-step stair climb	"… it certainly reflects, you know, about, uh, going up and down stairs at home. Uh, you know, because the —the height and the distance, uh, it—it mimics, you know, my everyday life I guess is what I'm trying to say." (R7, male, 57 y old, 58 d since surgery)	"I wouldn't go quickly unless sometimes the phone would ring and I would go up fast to answer it if I didn't take my phone downstairs with me. So, once in a great while I do go up quickly, but not generally." (F7, female, 81 y old, 171 d since fracture)
Long stair climb (eTHR)/ Repeated chair stand (HF)	"There was a little bit of difficulty at the end, but it was—it was reflective of what I—I go through at church and stuff, my daughter's house." (R1, female, 61 y old, 87 d since surgery)	"I usually have something in my hand or I'm—a book or whatever, you know? My phone or something in my hand. I usually don't have my arms crossed." (F16, female, 86 y old, 139 d since fracture)
	Patient views on the extent to which PerfO assessment level of difficulty reflects level of difficulty in everyday function	
Timed up and go	"… it's a little different because I have different thicknesses of carpeting, like throw rugs and stuff like that. It's not just the commercial carpet like where I had the test. The commercial carpet is easier because it doesn't have any give. It is just the floor and it is easier to walk on." (R8, female, 67 y old, 86 d since surgery)	"Probably more easy because I was just concentrating on the one thing. In life you get up to turn or you are thinking not of turning only, but of what you've got to do or why you're turning." (F3, female, 79 y old, 172 d since fracture)
Four-step stair climb	[The test was] "A little bit harder. I would say a little bit harder since there were more steps but not much." (R5, female, 72 y old, 89 d since surgery)	"It was similar except at the test I had two rails to hold on to. That made it a little bit easier. I only have the one at home. But I can push my hand against the other wall and hold on to the little handrail." (F8, female, 74 y old, 160 d since fracture)
Long stair climb (eTHR)/ Repeated chair stand (HF)	"I guess it was a good test for me to see if I would be able to do that in case I should have to. Like I said, I have not been doing this, but in case I should have to, in case of a fire or something like that, it was good to know that I could do it. I feel like I could. I would be able to do it." (R5, female, 72 y old, 89 d since surgery)	"Well, uh, if you're sitting on the couch watching TV and getting up, um, and that's on a soft and not a hard surface, uh, that was just as easy to me." (F2, female, 76 y old, 79 d since fracture)

eTHR, elective total hip replacement; HF, hip fracture; PerfO, performance outcome.

variations in specific aspects in daily life were described, such as chairs at home being lower or softer.

Four-step stair climb

All HF participants were able to relate the 4SC test directly to their regular experience of step-climbing, although two found it difficult to relate ascending the test staircase to when they climbed only a single step in daily life. There was considerable variation in the number of steps climbed in daily life from a single step to 16 steps.

Six of the eight eTHR participants were able to relate the 4SC test directly to the regular experience of climbing several steps that they had within their home. The two others were unable to think of instances in their daily life when they would routinely climb just a few steps at home or work, but could still relate the test to stepping up a curb or climbing one step into a shop.

Repeated chair stand
All HF participants were able to relate the RCS test to movements that they engaged in during their daily life, although five said that the frequent repetition during the test was not relevant to daily life; two reported that the version involving standing up from a chair with arms folded did not relate to any activity in their daily life.

Long stair climb
Most of the eTHR participants did not undertake longer stair climbs regularly in day-to-day life, but were nonetheless able to think of an example when they had climbed a set of longer stairs similar to the LSC, and reported that the test was an accurate reflection of their level of movement and ability.

Extent to Which Level of Difficulty in the PerfO Measures' Activities/Movements Reflects Level of Difficulty in Everyday Function

Example quotations from participants are presented in the third section of Table 4.

Timed up and go
All participants reported that overall the TUG test was feasible to complete. Four HF and five eTHR participants reported feeling that the level of difficulty was similar to that of related activities in daily life. Some participants mentioned that the TUG test was more difficult because of the feeling of pressure to complete the movement quickly, and the others reported ways in which the test was easier than everyday life, for instance, because of the shorter distance or easier walking surface.

Four-step stair climb
Most of the participants reported finding the 4SC test feasible to complete. Five eTHR participants reported that the level of difficulty in the 4SC test was comparable with climbing a few steps in daily life and that they felt the test reflected this well. Two others felt the test was easier, and one felt that it was slightly harder than steps used in daily life. Nine HF participants reported that the test was easier than steps in everyday life, and three participants reflected on how the test gave them confidence for considering or attempting more steps in daily life.

Repeated chair stand
All HF participants were able to complete the RCS test when using chair armrests, and 12 out of 18 participants were able to complete the RCS with their arms folded. Six participants reported that they felt that the level of difficulty between the test and daily life was similar. Five participants reported finding the test easier than similar movements in daily life, whereas two found the test harder. One participant reflected that the arms folded version was easier than anticipated, which gave him confidence to apply this in daily life.

Long stair climb
Most of the eTHR participants reported that the LSC test was a good reflection of the level of difficulty that they experienced when climbing a similar set of longer stairs in daily life. Nevertheless, most did not undertake longer stair climbs on a regular basis, with only two eTHR participants reporting the use of long staircases. Four eTHR participants reported climbing longer flights of stairs occasionally in their daily life, and two others reported that the test had shown them that they were able to climb a longer set of stairs.

Other daily activities affected by HF/eTHR and relationship to PerfO measure
Five HF participants and three eTHR participants mentioned daily activities that they felt indicated the level of difficulty or improvement associated with their hip that were not reflected in these PerfOs. These included getting in and out of vehicles (two HF and three eTHR participants), picking something up from the floor (one HF participant), and tying shoelaces (one HF participant). Participants felt that bending, leg lifting, and twisting movements of such activities were not captured by the PerfOs assessed.

Assessment of Data Saturation

A pragmatic approach to assessing saturation via participant summaries and review of new details was used for this study.

eTHR sample
In six of the nine core themes, new details were still identified in the final eTHR interview. These details, however, added variety without meaningful depth, reflected psychological considerations that digressed from study objectives, or reflected the specific PerfO measure instruction, for example, use of handrail for stability and not for pulling up. As a result, the new details identified in the final eTHR interview indicate that although additional interviews might have yielded further details, sufficient depth to understand the relevance and difficulty of the PerfO measures related to everyday life was achieved in the eight interviews conducted.

HF sample
New details were identified across all the nine core themes in the last three HF interviews. Nevertheless, these related to further variety in daily-life examples, responses, and thoughts about the PerfOs, for example, that the test was easy and so distance could be lengthened to make it more of a challenge. Therefore, the new details identified in the final HF interviews indicated that although additional interviews might have yielded further variety in details and additional broader areas of interest, they would be unlikely to add to the study objectives. This assessment suggests that sufficient depth of exploration and analysis was achieved from the 18 HF interviews.

Discussion

This study assessed the content validity of PerfO measures in participants who had undergone surgery for HF or eTHR. Unlike PRO measures that include assessment of how patients feel or function, this study focused on participants' perceptions of the activities and movements included in the PerfO measures and the relationship of those activities to functioning in their everyday lives. The three PerfO measures assessed for content validity in HF participants (TUG, 4SC, and RCS) were generally found to relate to activities and functioning in the participants' daily life, albeit with some variation in specific aspects. The three PerfO measures assessed for content validity in eTHR participants (TUG, 4SC, and LSC) revealed a range in their apparent correspondence to activities in everyday life: from all participants seeing similarity in the movements associated with the TUG test with everyday life to most undertaking short stair climbs in everyday life to most not undertaking longer stair climbs on a regular basis.

Innovative methods for assessing saturation were developed to meet the challenge of applying PRO guidance [5] to PerfO measures. Saturation was assessed by review of participant summaries, which indicated that a good depth and breadth in

experience was captured. Additional interviews might have provided further details; nevertheless, differences in this context could be considered as reflecting natural diversity in aspects of participant's mobility and daily-life contexts rather than lack of meaningful "saturation." Alternative approaches to explore the relevance of PerfO measures to functioning in daily life could include use of participant diaries or observation of the participant in their daily life (via wearable technology, for instance). Nevertheless, for the present study, such methods would have created additional burden on the participant, who had already agreed to participate in three main study visits, and additional interviews would still be needed to capture details that were specifically meaningful to the patient. Additional assessment of content validity could include triangulation with views of clinicians, physiotherapists, and close family members or friends who were familiar with the patients' abilities and daily activities (i.e., clinician- and observer-reported outcomes–assessing function).

Limitations of the study should be noted. First, there were some differences in sample characteristics compared with the main evaluation study populations; for example, the substudy sample comprised greater proportions of HF participants who used a staircase at home. Potentially substudy participants might have reported relatively better mobility, ability, and speed in PerfO measures and everyday life than the main study sample. Second, eTHR participants were interviewed after their third and final study visit. Had some interviews occurred after the presurgical main evaluation study visits, there would have been a greater range in responses: undertaking PerfO measure activities/movements would have been more difficult and the correspondence to activities completed in everyday life at that point could not be evaluated, although participants could reflect on this from memory. Third, the study did not routinely collect data about physiotherapy, which may have influenced participants' perception of their functioning. In addition, this study relied on participant recall during one-time telephone interviews that were conducted within 7 days of the PerfO visit. Interviews that were conducted sooner after PerfO measure completion may have reduced any recall bias. Finally, the sample size, particularly the eTHR subsample, was small. Data saturation assessment suggested that the eight interviews achieved sufficient depth to understand the relevance and difficulty of the PerfO measures for eTHR participants.

Methods and standards for PerfO measure development, selection, and implementation in study settings are evolving as more studies supporting clinical trials are conducted and regulatory feedback on industry plans is received. Currently, there is no regulatory guidance specific to PerfO measures, although FDA PRO guidance [5] provides a useful, albeit limited, framework. The establishment of measurement properties for PerfO measures presents specific challenges when compared with other COAs. Unlike for PRO measures, it is possible that some concepts of interest have less apparent direct correspondence to daily life from a patient's perspective (e.g., RCS with arms folded); nevertheless, the leg muscle strength required for this task may still have meaningful relevance to other, more usual, daily activities (e.g., standing up from a chair). While assessment of PRO measures focuses on domains and individual items, PerfO measures can be assessed in relation to specific components of movements (e.g., rising from a chair, turning, and stepping up). Each movement can be assessed in relation to how it corresponds to movement (and related level of difficulty) in daily life, for instance, how turning around a cone in the test differed from turning around in the kitchen at home. Nevertheless, interview guides need to be carefully structured to do so. Unlike PRO measure development where patient input can be used to craft items and create an instrument, patient input with PerfO measures could involve selection of measures and determining the extent to which measures reflect activities in daily life. Indeed some participants spontaneously reflected on activities they felt were relevant to their daily life that were not captured in the PerfO measures assessed. Future research could explore which movements are relevant in a patient's daily life to help in the selection of PerfO measures. It is, however, possible that had patients undergoing eTHR provided such input, assessments such as the LSC would not have been selected (because most participants did not regularly undertake these in daily life) or would have been included only at a lesser level of difficulty. The ability to see the positive impact of undertaking PerfO measures on activities in daily life would therefore have been missed in this validation study.

Conclusions

This study used a structured qualitative approach to assess the content validity of four PerfO measures in HF and eTHR participants enrolled in a main evaluation study assessing the measurement properties of PerfO measures. All HF and eTHR participants were able to ascertain relevance and level of difficulty of the PerfO measures in relationship to their everyday life, as well as recovery after surgery.

This study contributes novel methods in adapting the FDA's guidance for use of PRO measures in medical product development to support labeling claims for PerfO measure evaluation. A structured approach was used to explore the specific details of each measure and correspondence to everyday life, and data summary grids were developed to assess saturation. Even when there appears to be less apparent correspondence between aspects of PerfO measures to daily life and in the absence of guidance specific for PerfO measures, this study shows how content validity of PerfO measures can be meaningfully assessed.

Acknowledgments

We thank all the participants and study sites. We also thank Monique Curran for providing writing and editing assistance.

Source of financial support: This study was funded by Eli Lilly and Co.

REFERENCES

[1] US Food and Drug Administration. Clinical Outcome Assessment Qualification Program. Available from: http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm284077.htm. [Accessed February 25, 2015].

[2] Mizner R, Petterson S, Clements K, et al. Measuring functional improvement after total knee arthroplasty requires both performance-based and patient-report assessments. J Arthroplasty 2011;26:728–37.

[3] Kennedy DM, Stratford PW, Wessel J, et al. Assessing stability and change of four performance measures: a longitudinal study evaluating outcome following total hip and knee arthroplasty. BMC Musculoskelet Disord 2005;6:3.

[4] Dobson F, Hinman RS, Roos EM, et al. OARSI recommended performance-based tests to assess physical function in people diagnosed with hip or knee osteoarthritis. Osteoarthritis Cartilage 2013;21:1042–52.

[5] US Food and Drug Administration. US Food and Drug Administration Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. Rockville, MD: Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research, 2009.

[6] Walton MK, Powers JH, Hobart J, et al. Clinical outcome assessments: conceptual foundation—Report of the ISPOR Clinical Outcomes Assessment Emerging Good Practices for Outcomes Research Task Force. Value Health 2015;18:741–52.

[7] Rothman M, Burke L, Erickson P, et al. Use of existing patient-reported outcome (PRO) instruments and their modification: the

ISPOR Good Research Practices for Evaluating and Documenting Content Validity for the Use of Existing Instruments and Their Modification PRO Task Force Report. Value Health 2009;12: 1075–83.

[8] Doll H, Gentile B, Bush EN, et al. Evaluation of the measurement properties of four performance outcome measures in patients with elective hip replacements, elective knee replacements, or hip fractures. Value Health 2018, Forthcoming.

[9] Shumway-Cook A, Brauer S, Woollacott M. Predicting the probability for falls in community-dwelling older adults using the timed up and go test. Phys Ther 2000;80:896–903.

[10] Tiedemann A, Shimada H, Sherrington C, et al. The comparative ability of eight functional mobility tests for predicting falls in community-dwelling older people. Age Ageing 2008;37:430–5.

[11] Csuka M, McCarty DJ. Simple method for measurement of lower extremity muscle strength. Am J Med 1985;78:77–81.

[12] Kerr C, Nixon A, Wild D. Assessing and demonstrating data saturation in qualitative inquiry supporting patient-reported outcomes research. Expert Rev Pharmacoecon Outcomes Res 2010;10:269–81.

[13] Brod M, Tesler LE, Christensen TL. Qualitative research and content validity: developing best practices based on science and experience. Qual Life Res 2009;18:1263–78.