I SURVIVE: inter-rater reliability of three physical functional outcome measures in intensive care unit survivors

I SURVIVE : Fidélité interobservateurs de trois mesures de la capacité physique fonctionnelle chez des survivants de l'unité de soins intensifs

F. Aileen Costigan, OT Reg (Ont), PhD · Bram Rochweg, MD, MSc · Alexander J. Molloy, BSc ·
Magda McCaughan, MSc, PT · Tina Millen, RRT · Julie C. Reid, MSc, PT, PhD Candidate ·
Chris Farley, MSc, PT · Laurel Patterson, MSc, PT · Michelle E. Kho, PT, PhD

Received: 28 November 2018 / Revised: 4 March 2019 / Accepted: 22 March 2019 / Published online: 30 May 2019
© Canadian Anesthesiologists' Society 2019

Abstract

Purpose We prospectively assessed inter-rater reliability of three physical function measures in Canadian intensive care unit (ICU) survivors in the inpatient setting.

Methods We enrolled patients who had an ICU length of stay of $\geq$ three days, were mechanically ventilated for $\geq$ 24 hr, and were ambulating independently before hospital admission. Weekly from ICU discharge to hospital discharge, two trained frontline clinicians, blinded to each other's findings, independently performed the Physical Function ICU Test-scored (PFIT-s; score out of 10), 30-sec sit-to-stand (30STS; # of stands), and two-minute walk test (2MWT; distance in m), all within 24 hr. We calculated the intraclass correlation coefficient (ICC),

standard error of measurement (SEM), and minimal detectable change (MDC₉₀).

Results We enrolled 42 patients. PFIT-s: in 36 patients with 66 paired scores, the ICC was 0.78 (95% confidence interval [CI], 0.66 to 0.86), the SEM was 1.04, and the MDC₉₀ was 2.42. 30STS: in 35 patients with 67 paired scores, the ICC was 0.85 (95% CI, 0.76 to 0.90), the SEM was 1.91, and the MDC₉₀ was 4.45. 2MWT: in 35 patients with 58 paired scores, the ICC was 0.78 (95% CI, 0.66 to 0.87), the SEM was 20.87, and the MDC₉₀ was 48.69.

Conclusion These three measures show good inter-rater reliability when used by trained frontline clinicians to assess physical function in ICU survivors in the inpatient setting.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s12630-019-01411-x>) contains supplementary material, which is available to authorized users.

F. A. Costigan, OT Reg (Ont), PhD · A. J. Molloy, BSc ·
M. McCaughan, MSc, PT · L. Patterson, MSc, PT
Physiotherapy Department, St. Joseph's Healthcare, Hamilton,
ON, Canada

B. Rochweg, MD, MSc
Department of Medicine (Division of Critical Care), McMaster
University, Hamilton, ON, Canada

Department of Health Research Methods, Evidence and Impact,
McMaster University, Hamilton, ON, Canada

T. Millen, RRT · C. Farley, MSc, PT
Department of Critical Care, Juravinski Hospital, Hamilton, ON,
Canada

J. C. Reid, MSc, PT, PhD Candidate
Faculty of Health Sciences, School of Rehabilitation Science,
McMaster University, Hamilton, ON, Canada

M. E. Kho, PT, PhD (✉)
Physiotherapy Department, St. Joseph's Healthcare, Hamilton,
ON, Canada
e-mail: khome@mcmaster.ca

Faculty of Health Sciences, School of Rehabilitation Science,
McMaster University, Hamilton, ON, Canada

Institute of Applied Health Science, School of Rehabilitation
Science, McMaster University, Room 403, 1400 Main Street
West, Hamilton, ON L8S 1C7, Canada

Résumé

Objectif Nous avons évalué de manière prospective la fidélité interobservateurs de trois mesures de la capacité physique fonctionnelle auprès de survivants des unités de soins intensifs (USI) canadiennes dans un cadre hospitalier.

Méthode Nous avons recruté des patients qui avaient séjourné à l'USI $\geq$ trois jours, avaient été sous ventilation mécanique $\geq$ 24 h et se déplaçaient de façon autonome avant leur admission à l'hôpital. Entre le congé de l'USI et le congé de l'hôpital, deux cliniciens de première ligne formés ont réalisé chaque semaine et de manière indépendante trois tests : le PFIT (Test de la capacité physique fonctionnelle à l'USI - Physical Function ICU Test-scored; score sur 10), le test de 30 sec pour se lever d'une position assise (30STS; # de fois debout), et le test de marche pendant deux minutes (2MWT; distance en m). Les cliniciens n'avaient pas accès aux résultats de leur collègue. Nous avons calculé le coefficient de corrélation intraclasse (CCI), l'erreur type sur la mesure (ETM), et le changement minimal détectable (CMD₉₀).

Résultats Nous avons recruté 42 patients, et observé les résultats suivants : pour le PFIT : chez 36 patients avec 66 scores appariés, le CCI était de 0,78 (intervalle de confiance [IC] 95 %, 0,66 à 0,86), l'ETM de 1,04, et le CMD₉₀ de 2,42; pour le test 30STS : chez 35 patients avec 67 scores appariés, le CCI était de 0,85 (IC 95 %, 0,76 à 0,90), l'ETM de 1,91, et le CMD₉₀ de 4,45; pour le 2MWT : chez 35 patients avec 58 scores appariés, le CCI était de 0,78 (IC 95 %, 0,66 à 0,87), l'ETM de 20,87, et le CMD₉₀ de 48,69.

Conclusion Ces trois mesures affichent une bonne fidélité interobservateurs lorsqu'elles sont utilisées par des cliniciens de première ligne formés afin d'évaluer la capacité physique fonctionnelle des survivants de l'USI dans un cadre hospitalier.

Survivors of the intensive care unit (ICU) face important physical function and mobility impairments, which can last up to eight years post-ICU admission.¹ With more patients surviving the ICU,^{2,3} more are also at risk for post-ICU physical function impairments.^{1,4} Effective ICU-based interventions to maintain or improve patients' physical function post-ICU are urgently needed. In parallel, we need validated and rigorous measures to assess the effectiveness of these interventions in improving physical function in ICU survivors.

Despite the importance of measuring physical function in ICU survivors, there is no current consensus on the best instrument(s) to use.⁵ As part of the movement toward

establishing a core set of outcome measures, a recent Delphi study failed to reach consensus on any physical function outcome measures for studying patients with acute respiratory failure.⁶ One explanation for this lack of consensus may be that available measures of physical function lack comprehensive investigation of clinimetric properties, including reliability and responsiveness, and have not been validated in ICU survivors, particularly at different time points across the recovery trajectory.

A systematic review of 47 ICU studies found that only 12 of 26 functional outcomes included clinimetric properties and only six were developed explicitly for the ICU population.⁷ Thus, only a limited pool of physical function measures exist that have established any psychometric characteristics in ICU survivors, which is a unique population that may experience fatigue and difficulty with tasks that are typically included in these measures.⁵ Investigations of simple and brief one-item measures reporting the patient's highest level of performance (e.g., 30-sec sit-to-stand test [30STS],⁸ two-minute walk test [2MWT])^{9,10} may be especially useful in expanding the pool of physical function measures appropriate for ICU survivors.⁵

Previous investigations of physical function measures used in critically ill patients have not typically measured recovery beyond the ICU. For example, while investigations of the clinimetric properties of the Physical Function ICU Test - scored (PFIT-s) exist, they are confined to isolated time points early in the patient's recovery (e.g., ICU awakening, immediately post weaning, and ICU discharge).^{7,11} Given the potential for important change in patients' physical function post-ICU discharge,¹² investigations of the properties of appropriate outcome measures in ICU survivors across the hospital stay are needed. We prospectively analyzed the inter-rater reliability of three physical function measures, administered by trained frontline clinicians, in ICU survivors in the acute care setting. We report our findings according to the Guidelines for Reporting Reliability and Agreement Studies (eTable, available as Electronic Supplementary Material).¹³

Methods

Setting and study participants

Our study was approved by the Hamilton Integrated Research Ethics Board (HiREB #943, January, 2016; and #1647, November, 2016). All recruited patients provided written informed consent. We conducted a prospective, observational study with adult patients ($\geq$ 18 yr old) from two medical-surgical academic ICUs in Hamilton, ON,

Table 1 Inclusion and exclusion criteria

1	Inclusion criteria
1.1	Adult ≥ 18 yr old
1.2	Invasively mechanically ventilated ≥ 24 hr
1.3	ICU length of stay ≥ 3 days
1.4	Ability to ambulate independently pre-hospital (with or without gait aid)
1.5	Discharged from ICU or ready for ICU discharge within study hospital
1.6	Followed by physiotherapy in the ICU
1.7	First admission to ICU during this hospital stay
1.8	Likely to be available for 2 consecutive week 1 assessments
2	Exclusion criteria
2.1	Transferred to study ICU from another hospital with length of stay ≥ 7 days
2.2	Not walking ≥ 7 days during study hospital stay prior to ICU admission
2.3	Pre-hospital inability to follow simple commands in English at baseline
2.4	Acute proven or suspected central or peripheral neuromuscular weakness (e.g., stroke, Guillian Barré)
2.5	Other acute physical conditions impairing ability to perform assessments unresolved within 7 days of ICU discharge (e.g., leg fracture, unstable spinal injuries, pelvic fracture, skin integrity, etc.)
2.6	Other acute mental conditions impairing ability to perform assessments unresolved within 7 days of ICU discharge (e.g., delirium/not cognitively suitable)
2.7	Acute myocardial infarction with ongoing activity restrictions unresolved within 7 days of ICU discharge
2.8	Expected hospital mortality $\geq 90\%$
2.9	Palliative goals of care
2.10	Pregnancy

This table describes the inclusion and exclusion for I SURVIVE. Patients were required to meet all inclusion criteria and be free of exclusion criteria to be eligible for enrolment

ICU = intensive care unit

Canada. Table 1 outlines inclusion/exclusion criteria. These criteria were designed to recruit patients who were likely to experience physical function impairments post-ICU and who would also be able to participate in the outcome measure tests.¹⁴

Outcome measures

We included three measures to assess physical function (Table 2). The primary outcome measure was the four-task PFIT-s, a multi-faceted measure of physical function.^{15,16} We selected PFIT-s because several recent investigations established its basic clinimetric properties in ICU patients^{15,16} at ICU awakening and ICU discharge. Advantages of the PFIT-s are that it includes functional items and can be measured serially over time, in contrast to other measures such as the six-minute walk test (as few patients can walk at ICU awakening).¹⁷

Secondary outcome measures were the 30STS⁸ and 2MWT,^{9,10} both of which have age- and sex-based normative values, require no specialized equipment, are brief and simple to administer, and report the patient's highest level of performance such that ICU survivors should be able to participate despite physical limitations.

None of these three measures have specifically been studied in ICU survivors immediately following their ICU stay.

Study personnel

Twenty frontline clinicians across the two sites (18 physiotherapists and two physiotherapy assistants) assessed the outcome measures. All assessors received approximately 1.5 hr of training on the study protocol and how to administer the outcome measure tests from the methods centre (A.J.M., M.M., M.E.K.). All assessors were required to demonstrate competency by successfully administering the tests and scoring outcome measures on simulated patients as determined by the methods centre (A.J.M., M.M., M.E.K.) before participating in patient assessments. A pair of assessors, unaware of each other's scores on all outcome measures, completed a weekly assessment for each patient. These were performed until hospital discharge or death. Due to clinician schedules and patient availability, different pairs of assessors may have administered assessments for a particular patient from week to week.

Table 2 Description of outcome measures

Instrument	Brief description of measure	Scale	Measurement characteristics in ICU patients
PFIT-s ^{15,16}	The patient completes 4 tasks: sit-to-stand, marching on the spot, and muscle strength assessment (knee extension, shoulder flexion). Scores range from 0 to 10, with higher scores = better function	Continuous	Inter-rater reliability at ICU awakening and 48 hr post weaning from mechanical ventilation (ICC: time marching on the spot = 1.0 [95% CI, 0.99 to 1.00]; SEM= 0.762; not tested at hospital discharge in ICU population
30STS ⁸	The patient completes as many sit-to-stand movements as possible in 30 seconds, with higher number of repetitions = better strength	Continuous	Not tested in ICU population
2MWT ^{9, 10}	The patient walks as far as possible in 2 min, with farther distance in metres = better endurance	Continuous	Not tested in ICU population

This table describes the activities, measurement scale, and currently known clinimetric properties in ICU survivors of our three outcome measures

2MWT = two-minute walk test; 30STS = 30-sec sit-to-stand test; CI = confidence interval; ICC = intraclass correlation coefficient; MCID = minimal clinically important difference; PFIT-s = Physical Function ICU Test- scored; SEM = standard error of the measurement

Data collection

Research coordinators at each site identified and screened potential patients from a logbook that documented requests for patient transfer out of the ICU. For patients meeting inclusion criteria (Table 1) and with no exclusion criteria, the research coordinators contacted the medical and physiotherapy team at the time of ICU discharge to confirm that the patient was stable and appropriate for study enrolment. Once eligibility and stability were established, the research coordinators obtained informed consent for study participation from the patient or their substitute decision maker. As soon as possible after informed consent, the research coordinators arranged for two assessors to complete the baseline assessment based on clinician schedules and patient availability.

During baseline assessment, two assessors who were blinded to each other's findings administered all three outcome measures independently and in duplicate. Assessors prioritized the PFIT-s, our primary outcome, when a patient's capacity for completing all three assessments was considered questionable because of endurance or availability. Following baseline assessment, the research coordinators arranged weekly collection of the outcome measures, as described above, for each patient until hospital death or discharge. All pairs of assessments were administered by the assessors within a 24-hr time period of each other to minimize the effects of patient fatigue while maintaining measurement proximity.

We also collected baseline data including patient demographics, ICU admission reason, length of mechanical ventilation, length of stay (ICU, hospital), medical vs surgical status (defined as any operative procedure immediately preceding ICU admission), severity of illness,¹⁸ comorbidities,^{19,20} pre-hospital

function,²¹ highest level of mobility achieved in the ICU (e.g., sitting at the edge of the bed, walking), frailty,²² hospital discharge location, and vital status.

Sample size calculation

We determined that a sample size of 118 patients would be needed to determine whether an obtained inter-rater reliability of 0.7 is truly greater than a criterion standard given a minimally acceptable inter-rater reliability of 0.8 between two assessors, statistical power of 80%, and a one-tailed α of 0.05.^{23,24} After capturing the first 15 patients' paired PFIT-s scores, we observed a smaller than anticipated variation in paired scores and this led to a re-evaluation of our sample size calculation. We determined a sample size of 37 patients was sufficient given an expected intraclass correlation coefficient (ICC) of 0.88 and one-sided 95% confidence interval (CI) lower limit of 0.80 (one-tailed α of 0.05).

Statistical analysis

Pairs of assessments were eligible for inclusion in the analysis for each outcome if both assessors' assessments scored all components of the measure (i.e., for the PFIT-s, each of the four tasks received a numeric score from both assessors; for the 30STS, both assessors recorded the number of repetitions; for the 2MWT, both assessors recorded a distance in metres). We assigned a score of 0 when components of a score were missed because of pain, presence of intravenous lines, or if the assessor perceived the patient was unable to perform the component because of safety reasons. Since 0 is the lowest possible true score for all assessment components, we considered this to be an accurate representation of the patient's ability at the time of assessment.

We conducted all analyses in IBM SPSS Statistics for Windows, Version 25.0 (Armonk, NY, USA).²⁵ Descriptive statistics of continuous variables were calculated as means (standard deviations [SD]) or medians [interquartile ranges (IQR)] for skewed data. All descriptive data were visually inspected using boxplots and histograms. We used the Shapiro–Wilk test to determine normality. We calculated counts and percentages for categorical data. For each outcome, we determined inter-rater reliability, standard error of the measurement (SEM), and minimal detectable change (MDC₉₀). Inter-rater reliability was calculated by applying a one-way random (type 1,1) ICC^{26,27} for the total score of each measure. We calculated the SEM²⁷ to estimate measurement error in the

same units as the original measurement. We used the SEM to determine the minimal detectable change at a 90% confidence level (i.e., $SEM \times \sqrt{2} \times 1.645$).^{27,28}

Results

Figure 1 describes patient screening and enrolment. We enrolled 42 patients between October 5, 2016 and December 5, 2017. Our consent rate was 89% (42/47). Two patients died before participating in any outcome assessments; 40 patients ultimately participated in outcome measurement. Figure 2 describes the flow of patients through outcome assessments.

Fig. 1 Patient flow diagram. This figure describes patient screening and enrolment. * Patients may have had multiple reasons for exclusion or for non-enrolment despite eligibility such that total number of reasons may be greater than total number of patients with exclusions/eligible not enrolled. †Percent excluded patients was calculated as the proportion of patients with exclusion criteria compared with all those who met inclusion criteria (66/141). ‡Patient developmentally delayed. §Percent eligible but not enrolled was

calculated as proportion of eligible but not enrolled patients compared with the number of patients who met inclusion criteria and had no exclusion criteria (33/75). || Short-term hospital length of stay expected ($n = 1$); poor endurance/declining function ($n = 1$); history of non-compliance ($n = 1$). **CYCLE was a pilot randomized-controlled trial examining the effect of patient cycling in the ICU. These patients were excluded from the study because timing of the CYCLE outcome measures conflicted with the outcomes measured in the current study

Fig. 2 Patient flow through outcomes assessment. This figure describes flow through outcomes assessment to show the total number of patients who were included in reliability analyses for the PFIT-s, 30STS, and 2MWT. 30STS = 30-sec sit-to-stand test;

2MWT = two-minute walk test. ICU = intensive care unit; PFIT-s = Physical Function ICU Test- scored. *Medical issue with patient's visiting brother. †Patient needed a two-person assist to stand and could not stand to complete the task

Patient demographics

Patient characteristics are described in Table 3. Of the 40 included patients, 29 had a medical diagnosis (73%) and 11 (28%) had a surgical diagnosis. The median [IQR] Acute Physiology and Chronic Health Evaluation score was 20 [14–28].¹⁸ Before hospital admission, most patients were functioning independently with median [IQR] Katz Activities of Daily Living (i.e., a six-item measure of bathing, dressing, toileting, transferring, continence, and feeding, with higher scores indicating greater independence)²¹ and Functional Status Score for ICU (FSS-ICU)¹⁷ scores of 6 [6–6] and 42 [41–42], respectively. Most patients had maximum Katz (92%) and FSS-ICU (72%) scores, indicating high function in activities of daily living and mobility. The median [IQR] ICU length of stay was 7 [4–12] days, duration of mechanical ventilation 4 [2–9] days, and hospital length of stay 22 [16–48] days.

Physical function ICU test scored

Out of 40 patients, 36 (90%) had at least one pair of PFIT-s scores eligible for our reliability analysis. Median [IQR] number of paired scores per patient was 2 [1–2]. The minimum and maximum number of weeks a patient spent on study were one and eight, respectively. Across the 36

patients, we included 66 pairs of PFIT-s scores. Figure 3a shows a scatterplot of each pair of PFIT-s scores. The median [IQR] scores on score 1 and 2 were 7.90 [6.40–9.10] points and 8.80 [7.10–10.00] points, respectively. The estimated ICC was 0.78 (95% CI, 0.66 to 0.86). The SEM and the MDC₉₀ were 1.04 points and 2.42 points, respectively.

Thirty-second sit-to-stand

Out of 40 patients, 35 (88%) had at least one pair of 30STS scores eligible for analysis. Median [IQR] number of paired scores per patient was 2 [1–3]. Across the 35 patients, we included 67 pairs of 30STS scores. Figure 3b shows a scatterplot of each pair of 30STS scores. The median [IQR] score on assessments 1 and 2 were 5.00 [2.00–10.00] repetitions and 6.00 [4.00–10.00] repetitions, respectively. The estimated ICC was 0.85 (95% CI, 0.76 to 0.90). The SEM and the MDC₉₀ were 1.91 repetitions and 4.45 repetitions, respectively.

Two-minute walk test

Out of 40 patients, 35 (88%) had at least one pair of 2MWT scores eligible for analysis, and 58 pairs of scores. Median [IQR] number of paired scores per patient was 1 [1–2]. Figure 3c shows a scatterplot of each pair of 2MWT

Table 3 Patient demographics, baseline characteristics, and outcomes

Characteristic	Cohort (<i>n</i> = 40)
Age (yr), mean (SD)	62 (17)
Female, <i>n</i> (%)	21 (53%)
Race, <i>n</i> (%)	
White	38 (95%)
Aboriginal or African-Canadian	2 (6%)
APACHE II at ICU admission	20 [14–28]
Medical admission, <i>n</i> (%)	29 (73%)
Admission diagnosis, <i>n</i> (%)	
Gastrointestinal	13 (33%)
Respiratory	7 (18%)
Sepsis	6 (15%)
Neurologic	4 (10%)
Cardiovascular/vascular	4 (10%)
Renal	1 (3%)
Other	5 (13%)*
Pre-ICU Katz ADL score	6 [6–6]
Pre-ICU FSS-ICU	42 [41–42]
Location in hospital before ICU admission, <i>n</i> (%)	
Emergency room	19 (48%)
Hospital ward	7 (18%)
Operating/recovery room	10 (25%)
Surgical step-down unit	2 (5%)
Other hospital ICU or emergency room	2 (5%)
Duration of mechanical ventilation (days)	4 [2–9]
ICU length of stay (days)	7 [4–12]
Hospital length of stay (days)	22 [16–48]
Time from consent to 1 st assessment (days)	1 [0–1]
Time from ICU discharge to 1 st assessment (days)	3 [1–5]
Time from last assessment to hospital discharge/death (days)	3 [0–6]
Hospital mortality, <i>n</i> (%)	2 (5%)
Hospital discharge disposition of 38 survivors, <i>n</i> (%)	
Home, independent	10 (25%)
Home, with home care	10 (25%)
Home, with unpaid caregiver assistance	5 (13%)
Inpatient rehabilitation	5 (13%)
Other hospital	4 (10%)
Chronic care facility	2 (5%)

Table 3 continued

Characteristic	Cohort (<i>n</i> = 40)
Other	2 (5%)†

This table describes the demographics, baseline characteristics, and select patient outcomes. All measures are median [IQR] unless otherwise indicated

ADL = activities of daily living; APACHE II = Acute Physiology and Chronic Health Evaluation, a 13-item measure scored from 0 to 71 with higher scores indicating greater severity of illness;¹⁸ IQR = interquartile range; Katz ADL score is a six-item measure of bathing, dressing, toileting, transferring, continence, and feeding, with higher scores indicating greater independence;²¹ FSS-ICU = Functional Status Score for the Intensive Care Unit, a five-item performance-based functional measure including rolling in bed, sitting at the edge of the bed, sit-to-stand and bed-to-chair transfers, and ambulation. SD = standard deviation

*Other admission diagnoses included metabolic disease, necrotizing fasciitis, persistent atrial fibrillation and hypotension, and drug overdose. †Other discharge dispositions included acute inpatient mental health ward and alternative level of care ward

scores. The median [IQR] score on assessments 1 and 2 were 26.50 [0.00–66.25] m and 46.50 [6.25–82.00] m, respectively. The estimated ICC was 0.78 (95% CI, 0.66 to 0.87). The SEM and the MDC₉₀ were 20.87 m and 48.69 m, respectively.

Discussion

To effectively study interventions aimed at improving physical function in ICU survivors, it is imperative that measures are reliable, responsive, and valid during critical care recovery. Our prospective observational study enriches the existing literature by establishing key clinimetric properties of three additional outcome measures—the PFIT-s, the 30STS, and the 2MWT—in ICU survivors in the acute care setting. All three outcome measures showed good inter-rater reliability²⁹ with ICC values comparable with those recently reported for common physical function measures developed for the critically ill such as the ICU mobility scale (ICC = 0.80; 95% CI, 0.75 to 0.84) and the Chelsea Critical Care

Fig. 3 Scatterplots of score 1 vs score 2 for the PFIT-s, 30STS, and 2MWT. This figure shows the scatterplots of scores for the PFIT-s, 30STS, and 2MWT. Data points have been jittered to differentiate duplicate points. 2MWT = two-minute walk test; 30STS = 30-sec sit-to-stand test; PFIT-s = Physical Function ICU Test- scored

Physical Assessment Tool (CPAx) (ICC range = 0.97–1.00 for individual items).⁷

Investigations of the responsiveness of physical function outcome measures in the critically ill are exceptionally sparse. A recent systematic review found reports of responsiveness for only two physical function outcome measures for ICU patients: the PFIT-s, and the CPAx.⁷ In our study, all three outcome measures showed moderately low SEM in ICU survivors. Observed PFIT-s, 30STS, and 2MWT scores are likely within 1.04 points, 1.91 repetitions, and 20.87 m of a patient's true scores, respectively, and thus provide a reasonably accurate measurement of patients' actual physical abilities. Based on MDC₉₀ results, clinicians can be confident of a true change in physical function with differences of 2.42 points, 4.45 repetitions, and 48.69 m in PFIT-s, 30STS, and 2MWT scores, respectively. Our work expands the selection of physical function outcome measures that may be appropriate for monitoring patient recovery in the acute care setting after ICU discharge.

This is the first study to investigate the PFIT-s in Canadian critical care survivors' post-ICU discharge and

the first to investigate reliability of the total PFIT-s score. The ICC for the total PFIT-s score administered post-ICU is slightly lower than that established for the marching on the spot task administered in the ICU (Table 2).¹⁶ Previous investigations have only reported the SEM of individual PFIT-s tasks (Table 2),¹⁶ limiting direct comparison with our results for the total score. Denehy and colleagues reported a minimal clinically important difference for the PFIT-s of 1.5, the smallest difference in score that patients and/or clinicians are likely to consider important.¹⁵ In our study, we report the MDC₉₀ for the PFIT-s, which is the smallest change in score that does not arise from measurement error.²⁸

This study also investigates the clinimetric properties of both the 30STS and 2MWT in ICU survivors. Reliability and SEM of the 30STS in the critically ill were similar to the reliability obtained in recent studies with hospitalized older patients without ICU stays (ICC = 0.82, lower limit 95% CI = 0.67; SEM = 1.32)³⁰ and patients in an acute stroke unit (first session ICC = 0.88; 95% CI, 0.80 to 0.93; SEM = 0.95 m; second session ICC = 0.94; 95% CI, 0.90 to 0.96; SEM = 0.75 m).³¹ Reliability and SEM of the 2MWT

Table 4 Intraclass correlation coefficient scores for the PFIT-s, 30STS, and 2MWT

Outcome measure	<i>n</i>	# of paired scores	ICC (95% CI)	SEM	MDC ₉₀
PFIT-s	36	66	0.78 (0.66 to 0.86)	1.04	2.42
30STS	35	67	0.85 (0.76 to 0.90)	1.91	4.45
2MWT	35	58	0.78 (0.66 to 0.87)	20.87	48.69

This table describes the results of statistical analyses of reliability, standard error of the measure, and responsiveness for our three outcome measures

2MWT = two-minute walk test; 30STS = 30-sec sit-to-stand test; ICC = intraclass correlation coefficient; CI = confidence interval; MDC₉₀ = minimal detectable change at 90% confidence; PFIT-s = Physical Function ICU Test- scored; SEM = standard error of the measurement

in ICU survivors were similar to those found in a more recent study of community dwelling adults (ICC = 0.82; 95% CI, 0.76 to 0.87; SEM = 15.3 m); MDC at 95% confidence in this population was 42.5 m.⁹ Other recent investigations of the responsiveness of the 30STS and 2MWT have typically examined differences before and after intervention in patient groups with distinct differences from the critical care population (e.g., outpatients,³² patients undergoing joint replacements),³³ limiting comparisons with our results (Table 4).

Given their practicality and the reliability and responsiveness results of our study, the PFIT-s, 30STS, and 2MWT are strong candidates for use in future research on physical function in ICU survivors. Frontline acute care clinicians required minimal additional learning beyond clinical education and practice experience to effectively use these three tests to measure physical function in ICU survivors. Components of all three tests typically occur during an acute care physiotherapy treatment session. Thus, these can fit naturally into clinical practice with ICU survivors.

Future studies should include larger sample sizes and should address other relevant clinimetric properties of the PFIT-s, 30STS, and 2MWT including intra-rater reliability, concurrent validity with other relevant outcome measures (e.g., Functional Status Score for the Intensive Care Unit [FSS-ICU],¹⁷ CPAX),³⁴ and stability of reliability of estimates across the recovery trajectory. Future studies should also investigate the utility of the PFIT-s, the most comprehensive of our three outcome measures, in predicting other important outcomes for survivors of critical care (e.g., hospital discharge disposition), and in tracking change in patient functional status over time. Of particular importance are studies that establish the utility of the PFIT-s at key time points across the recovery trajectory and opportunities to transition to other outcome measures. Our study calculated overall reliability and responsiveness from pairs of assessment scores across an eight-week recovery period with most pairs of scores captured in week

1 (PFIT-s = 32/66 [48%], 30STS = 30/67 [45%], 2MWT = 28/58 [48%]).

Our results also show that the PFIT-s may be prone to ceiling effects and may not be sensitive to high-level changes in physical function that can occur later in the recovery period; 29% of all PFIT-s scores were maximum scores of 10 points and 18 of the 36 patients (50%) included in the PFIT-s analysis obtained at least one maximum score. The PFIT-s may be most useful earlier in the patients' recovery, with transition to the 30STS or 2MWT later on. Studies defining the time period over which the PFIT-s is most useful, along with investigations of other assessment tools that may serve as an alternative to the PFIT-s at particular time points across the recovery trajectory, are needed.

Limitations and strengths

We exceeded our target sample size by recruiting 40 patients in total. Nevertheless, only 36 patients with paired scores were eligible for the analyses of the PFIT-S and only 35 were eligible for both the 30STS and the 2MWT. These sample sizes were slightly under what was planned. Nevertheless, we included two medical-surgical ICUs to limit potential biases and included multiple pairs of scores per patient, when possible, to increase the number of score comparisons that inform our results and conclusions (PFIT-s = 66 pairs, 30STS = 67 pairs, 2MWT = 58 pairs). We scheduled all patient assessments according to the availability of the patient and the assessor, frequently negotiating scheduling challenges. As a result, assessors were not consistent across patients or across assessments (i.e., a particular assessor may have completed assessment 1 for one patient and assessment 2 for another patient). We are thus not able to control for the effect of the assessor or for the interaction between assessor and patient in our results. We addressed this issue, and the possibility that a single patient could contribute more than one pair of scores to our analyses, by using a one-way random (type 1,1) model, the most conservative option, to generate our

results. These results may underestimate the true reliability and responsiveness of the outcome measures.

Our study has several strengths. Assessors were frontline clinicians, professionals who are likely to be the ones to actually assess outcomes in clinical practice, thus enhancing the clinical applicability of the study. Paired assessments were conducted within a 24-hr period such that substantial changes in patient function were unlikely. Assessors were blind to each other's scores to limit the potential for expectation bias. Few assessment components were missed across outcome measures; the most common reason for a missing score was patient refusal. Despite complex clinician schedules, assessor availability did not often limit our ability to collect outcome measures.

Conclusions

It is feasible for frontline acute care clinicians with minimal additional training to reliably administer the PFIT-s, 30STS, and 2MWT to ICU survivors post-ICU discharge. True changes in patient functional status are likely indicated by score differences of 2.42 points, 4.45 repetitions, and 48.69 m for the PFIT-s, 30STS, and 2MWT respectively. The PFIT-s, 30STS, and 2MWT are reliable, responsive, and practical measures for measuring physical function in ICU survivors across the hospital stay and should be considered for integration into acute care clinical practice and research studies.

Acknowledgements The authors would like to thank the following people for their roles in this study: St. Joseph's Healthcare: Trained assessors: Sarah Couch, Caitlin Kuzyk, Frances Garner, Brittany McCulloch, Christine Neilson, Miranda Prince. Juravinksi Hospital: Trained assessors: Daniel Aranda, Cortney Armstrong, Mel Auclair, Helen Bishop, Tania Brittain, Rachel Consoli, Shivaun Davidson, Chelsea Hale, Jessie Koopman, Gillian Manson, Leigh Ann Niven, Kyla Soucie, Jessica Temesy; those who collected outcome measures or helped facilitate measurements; Lisa Buckingham for database support; Professor Paul Stratford and Dr. Lawrence Mbuagbaw for statistical consultation.

Declarations of interest None declared.

Editorial responsibility This submission was handled by Dr. Sangeeta Mehta, Associate Editor, *Canadian Journal of Anesthesia*.

Author contributions F Aileen Costigan, Bram Rochweg, and Michelle E. Kho contributed to all aspects of this manuscript, including study conception and design; acquisition, analysis, and interpretation of data; and drafting the article. Alexander J. Molloy contributed to acquisition, analysis, and interpretation of the data. Julie C. Reid contributed to the analysis and the interpretation of data. Magda McCaughan, Tina M. Millen, Chris Farley, and Laurel Patterson contributed to the acquisition of data.

Funding This work was supported by the Canadian Respiratory Health Professionals grant from the Lung Association and the Ontario

Ministry of Research, Innovation, and Science Early Researcher Award. Michelle Kho holds a Canada Research Chair in Critical Care Rehabilitation and Knowledge Translation from the Canadian Institutes of Health Research (CIHR). The funders had no influence on the design, conduct, analysis, or decision to submit this manuscript for publication.

References

1. Iwashyna TJ, Ely EW, Smith DM, Langa KM. Long-term cognitive impairment and functional disability among survivors of severe sepsis. *JAMA* 2010; 304: 1787-94.
2. Esteban A, Frutos-Vivar F, Muriel A, et al. Evolution of mortality over time in patients receiving mechanical ventilation. *Am J Respir Crit Care Med* 2013; 188: 220-30.
3. Kaukonen KM, Bailey M, Suzuki S, Pilcher D, Bellomo R. Mortality related to severe sepsis and septic shock among critically ill patients in Australia and New Zealand, 2000-2012. *JAMA* 2014; 311: 1308-16.
4. Needham DM, Davidson J, Cohen H, et al. Improving long-term outcomes after discharge from intensive care unit: report from a stakeholders' conference. *Crit Care Med* 2012; 40: 502-9.
5. Parry SM, Huang M, Needham DM. Evaluating physical functioning in critical care: considerations for clinical practice and research. *Crit Care* 2017; 21: 249.
6. Needham DM, Sepulveda KA, Dinglas VD, et al. Core outcome measures for clinical research in acute respiratory failure survivors. An International modified Delphi consensus study. *Am J Respir Crit Care Med* 2017; 196: 1122-30.
7. Parry SM, Granger CL, Berney S, et al. Assessment of impairment and activity limitations in the critically ill: a systematic review of measurement instruments and their clinimetric properties. *Intensive Care Med* 2015; 41: 744-62.
8. Jones CJ, Rikli RE, Beam WC. A 30-s chair-stand test as a measure of lower body strength in community-residing older adults. *Res Q Exerc Sport* 1999; 70: 113-9.
9. Bohannon RW, Wang YC, Gershon RC. Two-minute walk test performance by adults 18 to 85 years: normative values, reliability, and responsiveness. *Arch Phys Med Rehabil* 2015; 96: 472-7.
10. Pin TW. Psychometric properties of 2-minute walk test: a systematic review. *Arch Phys Med Rehabil* 2014; 95: 1759-75.
11. Reid JC, Kho ME, Stratford PW. Outcome measures in clinical practice: five questions to consider when assessing patient outcome. *Curr Phys Med Rehabil Rep* 2015; 3: 248-54.
12. Calvo-Ayala E, Khan BA, Farber MO, Ely EW, Boustani MA. Interventions to improve the physical function of ICU survivors: a systematic review. *Chest* 2013; 144: 1469-80.
13. Kottner J, Audigé L, Brorson S, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol* 2011; 64: 96-106.
14. Deney L, Skinner EH, Edbrooke L, et al. Exercise rehabilitation for patients with critical illness: a randomized controlled trial with 12 months of follow-up. *Crit Care* 2013; 17: R156.
15. Deney L, de Morton NA, Skinner EH, et al. A physical function test for use in the intensive care unit: validity, responsiveness, and predictive utility of the physical function ICU test (scored). *Phys Ther* 2013; 93: 1636-45.
16. Skinner EH, Berney S, Warrillow S, Deney L. Development of a physical function outcome measure (PFIT) and a pilot exercise training protocol for use in intensive care. *Crit Care Resusc* 2009; 11: 110-5.

17. Zanni JM, Korupolu R, Fan E, et al. Rehabilitation therapy and outcomes in acute respiratory failure: an observational pilot project. *J Crit Care* 2010; 25: 254-62.
18. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. *Crit Care Med* 1985; 13: 818-29.
19. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* 1987; 40: 373-83.
20. Groll DL, To T, Bombardier C, Wright JG. The development of a comorbidity index with physical function as the outcome. *J Clin Epidemiol* 2005; 58: 595-602.
21. Katz S, Ford AB, Moskowitz RW, Jackson BA, Jaffe MW. Studies of illness in the aged. the index of ADL: a standardized measure of biological and psychosocial function. *JAMA* 1963; 185: 914-9.
22. Bagshaw SM, Stelfox HT, McDermid RC, et al. Association between frailty and short- and long-term outcomes among critically ill patients: a multicentre prospective cohort study. *CMAJ* 2014; 186: E95-102.
23. Stratford PW, Spadoni G. Sample size estimation for the comparison of competing measures' reliability coefficients. *Physiotherapy Canada* 2003; 55: 225-9.
24. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. *Stat Med* 1998; 17: 101-10.
25. IBM. SPSS Statistics for Windows. 25.0th ed. Armonk, NY: IBM Corp; 2017 .
26. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979; 86: 420-8.
27. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res* 2005; 19: 231-40.
28. Riddle DL, Stratford P. *Is This Change Real?: Interpreting Patient Outcomes in Physical Therapy*. Paul Stratford, PA: FA Davis Co.; 2013 .
29. Koo TK, Li MY. A Guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016; 15: 155-63.
30. Bodilsen AC, Juul-Larsen HG, Petersen J, Beyer N, Andersen O, Bandholm T. Feasibility and inter-rater reliability of physical performance measures in acutely admitted older medical patients. *PloS One* 2015; 10: e0118248.
31. Lyders Johansen K, Derby Stistrup R, Skibdal Schjott C, Madsen J, Vinther A. Absolute and relative reliability of the timed 'up & go' test and '30second chair-stand' test in hospitalised patients with stroke. *PloS One* 2016; 11: e0165663.
32. Leung AS, Chan KK, Sykes K, Chan KS. Reliability, validity, and responsiveness of a 2-min walk test to assess exercise capacity of COPD patients. *Chest* 2006; 130: 119-25.
33. Unnanuntana A, Ruangsomboon P, Keesukpant W. Validity and responsiveness of the two-minute walk test for measuring functional recovery after total knee arthroplasty. *J Arthroplasty* 2018; 33: 1737-44.
34. Corner EJ, Wood H, Englebrechtsen C, et al. The Chelsea critical care physical assessment tool (CPAx): validation of an innovative new tool to measure physical morbidity in the general adult critical care population; an observational proof-of-concept pilot study. *Physiotherapy* 2013; 99: 33-41.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.