Published in final edited form as:
J Clin Epidemiol. 2017 February ; 82: 37–46. doi:10.1016/j.jclinepi.2016.08.014.

A systematic review finds limited data on measurement properties of instruments measuring outcomes in adult intensive care unit survivors

Karen A. Robinson, PhD[a,*], **Wesley E. Davis, BA**[b,e], **Victor D. Dinglas, MPH**[b,e], **Pedro A. Mendez-Tellez, MD**[c,e], **Anahita Rabiee, MD**[b,e], **Vineeth Sukrithan, MBBS**[b], **Ramakrishna Yalamanchilli, MBBS**[b], **Alison E. Turnbull, DVM, MPH, PhD**[b,d,e], and **Dale M. Needham, FCPA, MD, PhD**[b,e,f]

[a]Division of General Internal Medicine, Johns Hopkins University School of Medicine, 1830 East Monument St., Baltimore, MD, 21287

[b]Division of Pulmonary and Critical Care Medicine, Johns Hopkins University School of Medicine, 1830 East Monument St., Baltimore, MD, 21287

[c]Department of Anesthesiology and Critical Care Medicine, Johns Hopkins University School of Medicine, 600 N. Wolfe St., Baltimore, MD 21287

[d]Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, 615 North Wolfe St., Baltimore, MD

[e]Outcomes After Critical Illness and Surgery (OACIS) Group, Johns Hopkins University School of Medicine, 1830 East Monument St., Baltimore, MD, 21287

[f]Department of Physical Medicine and Rehabilitation, Johns Hopkins University School of Medicine, 600 North Wolfe St., Baltimore, MD, 21287

Abstract

Background and Objective—There is a growing number of studies evaluating the physical, cognitive, mental health and health-related quality of life (QOL) outcomes of adults surviving critical illness. However, there is little consensus on the most appropriate instruments to measure these outcomes. To inform the development of such consensus, we conducted a systematic review of the performance characteristics of instruments measuring physical, cognitive, mental health and HRQOL outcomes in adult intensive care unit (ICU) survivors.

Methods—We searched PubMed, Embase, PsycInfo, Cumulative Index of Nursing and Allied Health Literature, and The Cochrane Library in March 2015. We also conducted manual searches of reference lists of eligible studies and relevant review articles. Two people independently selected studies, completed data abstraction, and assessed the quality of eligible studies using the

[*]Corresponding author: Tel. +1-410-502-9216; krobin@jhmi.edu (K.A. Robinson).

Conflict of interest: None

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) initiative checklist.

Results—We identified 20 studies which explicitly evaluated measurement properties for 21 different instruments assessing outcomes in ICU survivors. Eleven of the instruments assessed QOL, with few instruments assessing other domains. Of the 9 measurement properties evaluated on the COSMIN checklist, 6 were assessed in <10% of the evaluations. Overall quality of eligible studies was poor to fair based on the COSMIN checklist.

Conclusions—While an increasing number of studies measure physical, cognitive, mental health and HRQOL outcomes in adult ICU survivors, data on the measurement properties of such instruments are sparse and generally of poor to fair quality. Empirical analyses evaluating the performance of instruments in adult ICU survivors are needed to advance research in this field.

Keywords

outcome measures; critical care survivors; systematic review

With the aging population leading to increased demand for critical care services, and with improving short-term mortality in the intensive care unit (ICU), there is a growing number of survivors of critical illness.[1, 2] Frequently such survivors experience significant challenges in their physical, cognitive, mental health and quality of life outcomes lasting long after hospital discharge.[3] Consequently, there are a growing number of studies evaluating post-discharge outcomes in adult ICU survivors.

More than 160 different outcome measures were identified in studies of adult ICU survivors in a 1998 systematic review.[4, 5] This systematic review reported on the validity, reliability and responsiveness of 38 instruments that had been used in at least two studies, but recommendations for selection of instruments were limited due to the poor quality of evidence. The great heterogeneity of measures and instruments led the authors to recommend the development of a limited set of outcome measures. A limited set of measures used in all future studies would improve comparability between studies and facilitate the synthesis of findings in this rapidly advancing field of research.

The heterogeneity of measures is not unique to critical care medicine. Across many fields of clinical research, there is growing interest in developing and adopting core outcome sets. Such “core outcome sets” outline a minimum set of measures to be reported in all studies of a particular health condition.[6] A key step in developing core outcome sets is understanding the measurement properties of outcome measurement instruments being considered for inclusion.

Hence, to help inform consensus and the development of a core outcome set[7] for evaluating post- discharge physical, cognitive, mental health and health-related quality of life (HRQOL) outcomes in adult survivors of critical illness[8], we conducted a systematic review of the measurement properties of instruments used in this population.

Methods

Our methods follow recommendations for conducting systematic reviews of measurement properties.[9, 10] In March 2015, we searched MEDLINE (via PubMed), Embase, Cumulative Index of Nursing and Allied Health Literature (CINAHL), PsycInfo, and The Cochrane Library (all databases, including Central Register of Controlled Studies, and Methodology Studies Database). We sought studies that reported or evaluated the measurement properties of instruments assessing health outcomes in survivors of intensive care. The search strategies combined controlled vocabulary and text words for intensive care and health outcomes, and were adapted from the strategy used in a prior related health technology assessment.[5] (See Appendix.) We also manually searched reference lists of eligible studies and relevant review articles identified by our search and by the COMET Initiative (Core Outcome Measures in Effectiveness Trials).[11] No limits were used for language, date or study design during the search phase.

Selection of Evidence

Two people independently screened search results. The pre-determined criteria for excluding studies were: (1) published prior to 1970, (2) non-English, (3) did not report the measurement properties of an instrument or test measuring physical, cognitive, mental health or quality of life outcome(s), (4) was not conducted in 20 or more adults (≥16 years) discharged from an ICU, or (5) only described instrument(s) that measured outcomes in-hospital (e.g., APACHE severity of illness score). We excluded studies prior to 1970 as we sought to focus on outcome measurement instruments currently in use and there is little research on ICU survivors prior to 1970.[12] We did not exclude studies based on type of health outcome assessed.

Data Abstraction and Assessment

Two people independently abstracted data including patient characteristics, sample size, timing of assessment, outcomes assessed, instruments used and measurement properties of the instruments.

We assessed the methodological quality of the studies that evaluated measurement properties using the 4-point rating scale checklist developed by the COnsensus-based Standards for the selection of health Measurement Instruments Initiative (COSMIN).[13] The COSMIN checklist contains 119 items across 12 boxes, nine of which assess the following measurement properties: internal consistency, reliability, measurement error, content validity, structural validity, hypothesis testing, cross-cultural validity, criterion validity and responsiveness. (See Table 1.) We judged the overall quality of how each measurement property was evaluated in an eligible study as excellent, good, fair, or poor per COSMIN methodology. Team members met several times prior to the start of the assessment to calibrate terminology and judgements and to pilot test the instrument. Thereafter, two people independently completed the COSMIN checklist for each study. Disagreement between researchers were resolved through discussion or adjudicated by a third person.

Data Synthesis

We classified each instrument by the type of outcome assessed using the three major domains within the International Classification of Functioning, Disability and Health (ICF) framework:[14] Body Functions and Structures (includes mental and physical impairments), Activity Limitation (includes cognitive and physical limitations), and Participation Restriction. As previously recommended, we added a domain for Health-Related Quality of Life to our classification.[15]

[13] Where more than one study assessed the measurement properties of an instrument we considered the methodological quality assessment (from COSMIN), consistency in results, and similarity of the studies in drawing overall conclusions.

Studies that did not explicitly assess measurement properties of instruments were categorized by type, but not synthesized because no hypotheses about the performance of the instruments were formulated or tested.

Results

We screened 18,647 citations, finding 39 articles reporting 37 studies (Figure 1), of which 20 met eligibility criteria for this systematic review. The remaining 17 studies did not directly assess measurement properties of an instrument and were thus ineligible for this analysis. (These studies were used only to develop an inventory of instruments. See Instrument Information http://www.improvelto.com/instruments/).

The 20 eligible studies examined 21 different instruments, with 85% of studies published after 1999. As shown in Table 2, we found very few studies across the different ICF domains. More than 50% of the identified instruments assessed quality of life, while we identified only one instrument measuring the ICF domain of Participation Restriction.

None of the studies evaluated all of the measurement properties considered by the COSMIN checklist (Table 3). For instance, while internal consistency and reliability comprised about 30% of the evaluations, another measure of the COSMIN domain reliability, ‘measurement error,’ was not assessed at all. Similarly, hypothesis testing (the degree to which scores on an instrument are correlated with objective measures or existing scales) comprised about 30% of the evaluations, but other properties of construct validity were assessed rarely (structural validity, 7% of evaluations) or not at all (cross-cultural validity). For the domain internal consistency, the most common reason for the rating of poor was a lack of factor analysis, while inadequate sample size, lack of details about comparator instrument and about the specific hypothesis being tested were the most common reasons for a rating of poor for each of the other COSMIN domains. Details of the studies are provided in Table 4 and study level summary of COSMIN scores is provided in a supplemental table (Summary of COSMIN Scores Supplement Table).

ICF: Body Functions and Structures

Anxiety and Depression—One study compared the Hospital Anxiety and Depression Scale (HADS) with the Depression Anxiety and Stress Scale (DAAS).[16] Both instruments were administered to ICU survivors at 3 and 9 months post-ICU, with the HADS averaging less time to complete (4 versus 10 minutes). The internal consistency was ‘good’ for DASS (Cronbach’s alpha 0.92 to 0.95) and for HADS (0.82 to 0.86); however, the overall COSMIN quality assessment for this evaluation was poor due to lack of factor analyses or item response theory (IRT) analyses. (Supplement Table). In terms of criterion validity, the reported correlations were strong between the two measures of anxiety (r=0.88, p<0.0001) and between the two measures of depression (r=0.93, p<0.0001), with a COSMIN quality rating of fair.

Post-traumatic Stress Disorder (PTSD)—Three studies evaluated four different PTSD instruments.[17-19] Investigators in Germany developed a tool that included 4 yes/no questions to assess traumatic memories as well as a modified German version of the Post-Traumatic Stress Syndrome 10-questions Inventory (PTSS-10) instrument.[17] The questions in the PTSS do not specifically link PTSD symptoms to an event (e.g., critical illness). Post-ICU and at a median of 2-years later, internal consistency for the PTSS- 10 was high (Cronbach’s alpha 0.91 and 0.93). Reliability over the 2 year time interval was also high (intraclass correlation coefficient (ICC)=0.89). However, the evaluation of reliability and internal consistency were considered poor per COSMIN assessment. This was due to a long (2-year) time interval between surveys and not completing factor analyses or IRT analyses, respectively.

In contrast to the PTSS, the IES-R is anchored to a specific event (e.g., critical illness). The IES-R instrument was compared to concurrent administration of a “gold standard” semi-structured interview, the Clinician-Administered PTSD Scale (CAPS).[19] The IES-R vs. CAPS takes less time (6 versus 60 minutes), and can be self-administered rather than needing a trained evaluator. The evaluation of the internal consistency of both instruments was considered poor per COSMIN because factor analyses or IRT analyses were not completed and internal consistency was not evaluated on the subscales. There was high correlation (r=0.80, Spearman p=0.69) between the IES-R total score and the CAPS total severity score (COSMIN assessment of criterion validity: fair).

Twigg (2008) extended the PTSS-10 to 14 questions by adding questions about numbing and flashbacks to become the UK PTSS-14.[18] Internal consistency was evaluated at 2 months (Cronbach’s alpha 0.86) and 3 months (0.84), but was considered poor per COSMIN due to not completing factor analyses or IRT analyses. Reliability of PTSS-14, also considered poorly evaluated per COSMIN (due to different testing conditions for each time point: hospital ward (2-14 days after ICU discharge), over phone (2 months after ICU discharge), and then in-person at follow-up clinic (3 months after discharge)), demonstrated ICC ranging from 0.70 to 0.90 across the two assessment time points. Criterion validity (COSMIN assessment: fair) was evaluated using receiver operating curve (ROC) analysis, with an area under the curve of 0.94 (95% confidence interval (CI) 0.82 to 0.99) at 3-months (last follow-up). Two measures previously validated in other patient populations – the Posttraumatic Stress Diagnostic Scale (PDS) and the Impact of Events Scale (IES) -- were used as in assessing the predictive ability of PTSS-14. At the 3 month assessment, Pearson’s correlation with IES was 0.71 (95% confidence interval (CI) 0.52 to 0.83) and with PDS was 0.85 (95% CI 0.74 to 0.92) (COSMIN quality assessment: fair).

Physical and Mental Impairment—We identified two studies by the same investigators reporting the development and subsequent assessment of the 3-set 4P instrument (4P refers to Patients' Physical and Psychosocial Problems).[20, 21] This 53-item instrument assesses disability and need for follow-up care after the ICU. In the development paper, the evaluation of internal consistency, reliability and structural validity had a COSMIN quality assessment of poor due to a small sample size, while the second paper further evaluated the same measurement properties in a larger population with a COSMIN quality assessment of fair.

ICF: Activity Limitation

Cognitive Limitations—Christie (2006) assessed reliability (COSMIN rating of fair), content validity (poor) and hypothesis testing (fair) of a battery of tests developed by combining elements from standardized tools (the Cognitive Telephone Battery).[22] The telephone administration took 20 to 30 minutes to complete.

Physical Limitations—Two studies evaluated the 6-minute walk test (6MWT), examining expected correlations with other instruments.[23, 24] Alison (2012) assessed the relationship between the 6MWT and the physical functioning score (PF) of the SF-36, reporting a moderate to high correlation across measurements at 1, 8 and 26 weeks post-ICU (r= 0.62, 0.55. 0.47, all p<0.001; COSMIN assessment: poor due to an unclear hypothesis for the correlation analysis).[24] In the other study, Denehy and colleagues (2014) also used the SF-36 PF for comparison (Spearman's rho=0.69), as well as the physical component score (PCS) of the SF-36 (rho=0.50), reporting moderate correlation (COSMIN quality assessment: fair).[23] Given that only one measurement property was assessed we are unable to draw overall conclusions about the use of this instrument in ICU survivors.

ICF: Participation Restriction

Participation in Usual Environment—We identified only one study assessing the measurement properties of an instrument measuring functional status.[25] Jones and colleagues (1993) developed a test of health status in survivors (Follow-up Health Status Questionnaire (FHS)) by modifying the Pre-Morbid Health Status (PHS) questionnaire, and comparing with the Functional Limitations Profile (FLP) and Perceived Quality of Life instrument (PQL). The FHS includes items related to employment status, current living location, and quality of life. They reported high correlation with FLP (r=0.7, p<0.001) and with PQL (r=0.678, p<0.001), with a poor COSMIN quality assessment due to not reporting measurement properties of the comparator instruments.

Health-Related Quality of Life

Skinner et al. (2013) compared measurement properties of the Assessment of Quality of Life (AQOL) and the Medical Outcomes Study Short Form 6D (SF-6D).[26] Internal consistency was moderate for AQOL (Cronbach's alpha=0.81) and lower for the SF-6D (Cronbach's alpha=0.65), with a poor COSMIN quality rating due to not completing factor analyses or IRT analyses. There was moderate agreement between the scales (COSMIN quality assessment: fair).

Consisting of only 5 questions, the EQ-5D takes substantially less time for patients to complete than either the SF-36 or RAND-36. Three studies assessed the EQ-5D: Kaarlola and colleagues (2004) compared the EQ-5D with the RAND-36;[27] Khoudri and colleagues assessed reliability of the EQ-5D, also comparing it with the SF-36;[28, 29] and Vainiola et al. (2010) compared the EQ-5D with the 15D HRQOL [30] However, while three studies conducted evaluations of the EQ-5D, the specific measurement properties assessed differed. Kaarlola (2004) conducted hypothesis testing (COSMIN quality assessment: fair) finding strong correlation between EQ-5D and SF-36 v1. Khoudri and colleagues assessed reliability for EQ-5D, as well as for the subscales of the instrument, with COSMIN quality of the evaluation being considered fair to good. Criterion validity evaluation was judged to be excellent. Vainiola and colleagues used the EQ-5D as comparison for the 15D Health Related Quality of Life (HRQOL) instrument for hypothesis testing (COSMIN quality assessment: fair) and responsiveness (COSMIN quality assessment: poor, due to not reporting measurement properties of the comparator instrument).

Fernandez et al. (1996) developed a 15-item questionnaire with subscales assessing basic physiological activities, normal daily activities, and emotional state.[31] Internal consistency (Cronbach's alpha = 0.85), reliability (Spearman Rho=0.92), and hypothesis testing (concordance with Glasgow Outcome Scale, p<0.01 Kruskal Wallis test) were assessed across the whole questionnaire as well as the subscales (COSMIN quality assessment: fair). Structural validity was considered for the three subscales (factor loadings 0.52 to 0.84) (COSMIN quality assessment: fair) and responsiveness (weighted kappa with Glasgow Outcome Scale =0.56) was assessed for the whole questionnaire (COSMIN quality assessment: fair).

Capuzzo et al. (2000) assessed two questionnaires developed in 1996 for use in Italy: the Spanish questionnaire developed by Fernandez and colleagues, and an Italian questionnaire developed by Capuzzo and colleagues. [32] Their assessments of the internal consistency, reliability, hypothesis testing and responsiveness for each instrument was considered poor to fair via COSMIN checklist. Both scales were reported with moderate internal consistency (Cronbach's alpha: QOL-IT 0.76, QOL-SP 0.84), and good reliability (QOL-IT kappa>0.90, QOL-SP kappa>0.90). Both were also determined to be responsive in comparing patients with and without functional limitations (Wilcoxon test p-value <0.01).

The Modified Short-Form-36 (MSF-36), which shortens the SF-36 to 20 questions from 36, was assessed at 1, 3 6, and 12 months by Lipsett et al. (2000).[33] Internal consistency was poor at 1 and 3 months (Cronbach's alpha 0.59 and 0.35), but was moderate to high at 6 and 12 months (Cronbach's alpha 0.87 and 0.93). The reliability at 12 months was good (r >0.92), although this was conducted in only 10 of the 127 patients. The Sickness Impact Profile (SIP) was also assessed, with higher internal consistency (Cronbach's alpha 0.92 to 0.95 across time points). All internal consistency assessments had poor COSMIN quality assessments. The MSF-36 was found to correlate moderately well with SIP at 12 months (COSMIN quality assessment: fair).

One study assessed the Satisfaction Scale, SIP, and a 'standardized telephone interview'.[34] Frick et al. (2002) used SIP to assess the criterion validity of the Satisfaction Scale and the

Telephone Interview (COSMIN assessment for both instruments: fair). SIP compared with the Satisfaction Scale showed good correlation (COSMIN quality assessment: fair).

Two studies assessed the SF-36 version 1 (SF-36 v1). Khoudri and colleagues included an examination of internal consistency (COSMIN assessment: poor) and reliability of subscales of SF-36 v1 (fair).[28] They reported good internal consistency and reliability for each subscale. Heyland et al. (2000) also reported good internal consistency (COSMIN assessment: poor due to not completing factor analyses or IRT analyses) and reliability (COSMIN quality assessment: fair) across the subscales of SF-36 v1.[35]

Discussion

Choosing appropriate measures to assess outcomes in research evaluating ICU survivors is essential, but challenging. The development of core outcome sets, through consideration of studies evaluating the measurement properties of outcome measures, as well as expert and patient/caregiver input, would help addresses this challenge. The importance of this work is highlighted by the work of COMET, as well as by the Outcome Measures Working Group of InFACT (International Forum for Acute Care Trialists).[36]

We conducted a systematic review of studies of measurement properties of instruments evaluating >8,500 adult ICU survivors; a needed step to understanding the performance of existing outcome measures as part of the development of a core outcome set. We identified 20 studies that explicitly assessed the measurement properties of 21 instruments. Similar to the prior review of instruments that had been used in at least two studies,[4, 5] we were unable to draw conclusions about the quality of any specific outcome measurement instrument. This suggests that there has been little to no improvement over the past 18 years.

Our classification of identified instruments found very few studies evaluating measures for the different ICF domains. More than 50% of the studies evaluated instruments measuring health-related quality of life; only one instrument assessed the ICF domain of Participation Restriction. It is possible that gaps would be even more apparent if evaluated using an outcome framework specific to critical care survivorship. The development of such a conceptual framework, incorporating and moving beyond the ICF classification, would be useful in developing core outcome sets for research in critical care medicine.[7]

The overall quality of the studies, as assessed by the COSMIN checklist, was generally poor or fair. Inadequate sample size, and failure to report details about the comparator (i.e., reference or other instrument) and hypothesis being tested, were the most frequent reasons for these COSMIN ratings. The latter quality limitation may be readily improved via more detailed reporting in future studies. In addition to the general poor to fair quality, several measurement properties were rarely or never assessed. Of the 9 properties assessed by the COSMIN checklist, 6 were assessed in fewer than 10% of the evaluations.

However, as with the critical appraisal of other types of studies, we are limited in assessing the quality of the studies to assessing what was reported in the articles. Accentuating this problem is the fact that standards for reporting studies of measurement properties do not exist. Such standards would aid those conducting systematic reviews, and, more generally, help to ensure that the reports of such studies are maximally informative. The COSMIN checklist may assist with such considerations in future studies.

Another reporting issue is the heterogeneity in terms and definitions of performance characteristics used in the literature. To the extent possible, we matched the characteristic being assessed in the study with the terminology used in the COSMIN checklist, based on the definition provided in the study. We recommend that researchers conducting studies of instruments should review elements in COSMIN to ensure complete reporting of study details needed for interpretation.

The development and evaluation of instruments would be also be improved through greater emphasis on research in measurement property evaluation, including a focus on training and encouragement of collaboration. Specifically, we suggest that those with domain expertise should work with someone with expertise in assessing measurement properties in developing and evaluating instruments. This sort of work may be supported through mechanisms that support research projects to develop and enhance resources for efficient and effective research. For instance, this review was part of a larger project funded via the NIH R24 mechanism; the larger project includes international collaboration on new analyses of outcome measurement properties (see www.improvelto.com). Further, this review will serve as a basis for a systematic process for developing a core outcome measurement set following the process developed by the OMERACT (Outcome Measures in Rheumatology) initiative.[37]

The international efforts for standardization of outcome measures will also place a greater emphasis on research in outcome measurement. These efforts, including COMET and, specifically in critical care, InFACT, demand rigorous examination of outcome measurement and will move the field forward. As reviews are completed (or updated) to inform the development of outcome measure sets, we will see if these efforts to improve the field of instrument development and evaluation have made a difference.

A lack of studies, poorly conducted studies, and poorly reported studies meant there was insufficient evidence to draw conclusions about the measurement quality of any of the instruments that measure physical, cognitive, mental health or quality of life outcomes in adult ICU survivors. There is an urgent need for empirical analyses evaluating the performance of instruments and tests in adult survivors of critical illness to advance research in this field.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding: This project was supported through a grant from the National Heart, Lung and Blood Institute (NHLBI R24HL111895).

Appendix: PubMed Search Strategy

(intensive care[tiab] OR "intensive care"[MeSH Terms] OR intensive therapy[tiab] OR high dependency[tiab] OR critical care[tiab] OR "critical care"[MeSH Terms] OR intermediate care[tiab] OR step-up care[tiab] OR step-down care[tiab] OR respiratory distress syndrome[tiab] OR acute lung injury[tiab]) AND (outcome measure[tiab] OR "outcome assessment (health care)"[MeSH Terms] OR follow-up[tiab] OR "follow-up studies"[MeSH Terms] OR health status[tiab] OR "health status"[MeSH Terms] OR functional status[tiab] OR clinical outcome[tiab]) AND (organ failure[tiab] OR "multiple organ failure"[MeSH Terms] OR organ dysfunction[tiab] OR sequelae[tiab] OR quality of life[tiab] OR "quality of life"[MeSH Terms] OR impairment[tiab] OR morbidity[tiab] OR "morbidity"[MeSH Terms]) NOT (animals[mh] NOT humans[mh])

References

1. Spragg RG, Bernard GR, Checkley W, et al. Beyond mortality: future clinical research in acute lung injury. American journal of respiratory and critical care medicine. 2010 May 15; 181(10):1121–7. [PubMed: 20224063]
2. Needham DM, Bronskill SE, Calinawan JR, et al. Projected incidence of mechanical ventilation in Ontario to 2026: Preparing for the aging baby boomers. Critical Care Medicine. 2005 Mar; 33(3): 574–9. [PubMed: 15753749]
3. Desai SV, Law TJ, Needham DM. Long-term complications of critical care. Critical Care Medicine. 2011 Feb; 39(2):371–9. [PubMed: 20959786]
4. Black NA, Jenkinson C, Hayes JA, et al. Review of outcome measures used in adult critical care. Critical Care Medicine. 2001 Nov; 29(11):2119–24. [PubMed: 11700407]
5. Hayes JA, Black NA, Jenkinson C, et al. Outcome measures for adult critical care: a systematic review. Health technology assessment. 2000; 4(24):1–111.
6. Gargon E, Gurung B, Medley N, et al. Choosing important health outcomes for comparative effectiveness research: a systematic review. PloS one. 2014; 9(6):e99111. [PubMed: 24932522]
7. Blackwood B, Marshall J, Rose L. Progress on core outcome sets for critical care research. Current opinion in critical care. 2015 Aug 8.
8. Needham DM, Davidson J, Cohen H, et al. Improving long-term outcomes after discharge from intensive care unit: report from a stakeholders' conference. Critical Care Medicine. 2012 Feb; 40(2): 502–9. [PubMed: 21946660]
9. de Vet, HCW., Terwee, CB., Mokkimk, LB., et al. Measurement in Medicine: A Practical Guide. Cambridge: Cambridge University Press; 2011.
10. Mokkink LB, Terwee CB, Stratford PW, et al. Evaluation of the methodological quality of systematic reviews of health status measurement instruments. Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation. 2009 Apr; 18(3):313–33.
11. Terwee, CB. group ftC. Amsterdam: 2014. Systematic reviews of measurement instruments that have used the COSMIN checklist. http://www.cosmin.nl/images/upload/files/Systematic%20reviews%20using%20COSMIN.Pdf [22 July 2015]
12. Turnbull AE, Rabiee A, Davis WE, et al. Outcome Measurement in ICU Survivorship Research From 1970 to 2013: A Scoping Review of 425 Publications. Critical Care Medicine. 2016 May 17.
13. Terwee CB, Mokkink LB, Knol DL, et al. Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation. 2012 May; 21(4):651–7.
14. International Classification of Functioning, Disability and Health (ICF). Geneva: World Health Organization; 2001. http://www.who.int/classifications/icf/en/ [17 July 2015]

15. Iwashyna TJ, Netzer G. The burdens of survivorship: an approach to thinking about long-term outcomes after critical illness. Seminars in respiratory and critical care medicine. 2012 Aug; 33(4): 327–38. [PubMed: 22875378]

16. Sukantarat KT, Williamson RC, Brett SJ. Psychological assessment of ICU survivors: a comparison between the Hospital Anxiety and Depression scale and the Depression, Anxiety and Stress scale. Anaesthesia. 2007; 62(3):239–43. [PubMed: 17300300]

17. Stoll C, Kapfhammer HP, Rothenhausler HB, et al. Sensitivity and specificity of a screening test to document traumatic experiences and to diagnose post-traumatic stress disorder in ARDS patients after intensive care treatment. Intensive Care Med. 1999; 25(7):697–704. [PubMed: 10470573]

18. Twigg E, Humphris G, Jones C, et al. Use of a screening questionnaire for post-traumatic stress disorder (PTSD) on a sample of UK ICU patients. Acta Anaesthesiol Scand. 2008; (2):202–8. 2007/11/17. [PubMed: 18005373]

19. Bienvenu OJ, Williams JB, Yang A, et al. Posttraumatic stress disorder in survivors of acute lung injury: evaluating the Impact of Event Scale-Revised. 2013; 144(1):24–31.

20. Akerman E, Fridlund B, Ersson A, et al. Development of the 3-SET 4P questionnaire for evaluating former ICU patients' physical and psychosocial problems over time: a pilot study. Intensive Crit Care Nurs. 2009; 25(2):80–9. [PubMed: 18692395]

21. Akerman E, Fridlund B, Samuelson K, et al. Psychometric evaluation of 3-set 4P questionnaire. Intensive Crit Care Nurs. 2013; 29(1):40–7. [PubMed: 22835992]

22. Christie JD, Biester RC, Taichman DB, et al. Formation and validation of a telephone battery to assess cognitive function in acute respiratory distress syndrome survivors. J Crit Care. 2006; (2): 125–32. 2006/06/14. [PubMed: 16769455]

23. Denehy L, Nordon-Craft A, Edbrooke L, et al. Outcome measures report different aspects of patient function three months following critical care. Intensive Care Med. 2014; 40(12):1862–9. [PubMed: 25319384]

24. Alison JA, Kenny P, King MT, et al. Repeatability of the six-minute walk test and relation to physical function in survivors of a critical illness. Phys Ther. 2012; 92(12):1556–63. [PubMed: 22577064]

25. Jones C, Hussey R, Griffiths RD. A tool to measure the change in health status of selected adult patients before and after intensive care. Clin Intensive Care. 1993; (4):160–5. 1992/12/09. [PubMed: 10146456]

26. Skinner EH, Denehy L, Warrillow S, et al. Comparison of the measurement properties of the AQoL and SF-6D in critical illness. Critical care and resuscitation : journal of the Australasian Academy of Critical Care Medicine. 2013 Sep; 15(3):205–12. [PubMed: 23944207]

27. Kaarlola A, Pettila V, Kekki P. Performance of two measures of general health-related quality of life, the EQ-5D and the RAND-36 among critically ill patients. Intensive Care Med. 2004; (12): 2245–52. 2005/01/15. [PubMed: 15650867]

28. Khoudri I, Ali Zeggwagh A, Abidi K, et al. Measurement properties of the short form 36 and health- related quality of life after intensive care in Morocco. Acta Anaesthesiol Scand. 2007; (2): 189–97. 2007/01/31. [PubMed: 17261146]

29. Khoudri I, Belayachi J, Dendane T, et al. Measuring quality of life after intensive care using the Arabic version for Morocco of the EuroQol 5 Dimensions. BMC Res Notes. 2012:56. 2012/01/24. [PubMed: 22264312]

30. Vainiola T, PettilA V, Roine RP, et al. Comparison of two utility instruments, the EQ-5D and the 15D, in the critical care setting. Intensive Care Medicine. 2010; 36(12):2090–3. [PubMed: 20689933]

31. Fernandez RR, Cruz JJ, Mata GV. Validation of a quality of life questionnaire for critically ill patients. Intensive Care Med. 1996; 22(10):1034–42. [PubMed: 8923066]

32. Capuzzo M, Grasselli C, Carrer S, et al. Validation of two quality of life questionnaires suitable for intensive care patients. Intensive Care Med. 2000; (9):1296–303. 2000/11/23. [PubMed: 11089756]

33. Lipsett PA, Swoboda SM, Campbell KA, et al. Sickness Impact Profile Score versus a Modified Short-Form survey for functional outcome assessment: acceptability, reliability, and validity in

critically ill patients with prolonged intensive care unit stays. J Trauma. 2000; (4):737–43. 2000/10/19. [PubMed: 11038094]

34. Frick S, Uehlinger DE, Zurcher Zenklusen RM. Assessment of former ICU patients' quality of life: comparison of different quality-of-life measures. Intensive Care Med. 2002; (10):1405–10. 2002/10/10. [PubMed: 12373464]
35. Heyland DK, Hopman W, Coo H, et al. Long-term health-related quality of life in survivors of sepsis. Short Form 36 A valid and reliable measure of health-related quality of life. Critical Care Medicine. 2000; 28(11):3599–605. [PubMed: 11098960]
36. Shea BJ, Grimshaw JM, Wells GA, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC medical research methodology. 2007; 7:10. [PubMed: 17302989]
37. Boers M, Kirwan JR, Wells G, et al. Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. Journal of clinical epidemiology. 2014 Jul; 67(7):745–53. [PubMed: 24582946]

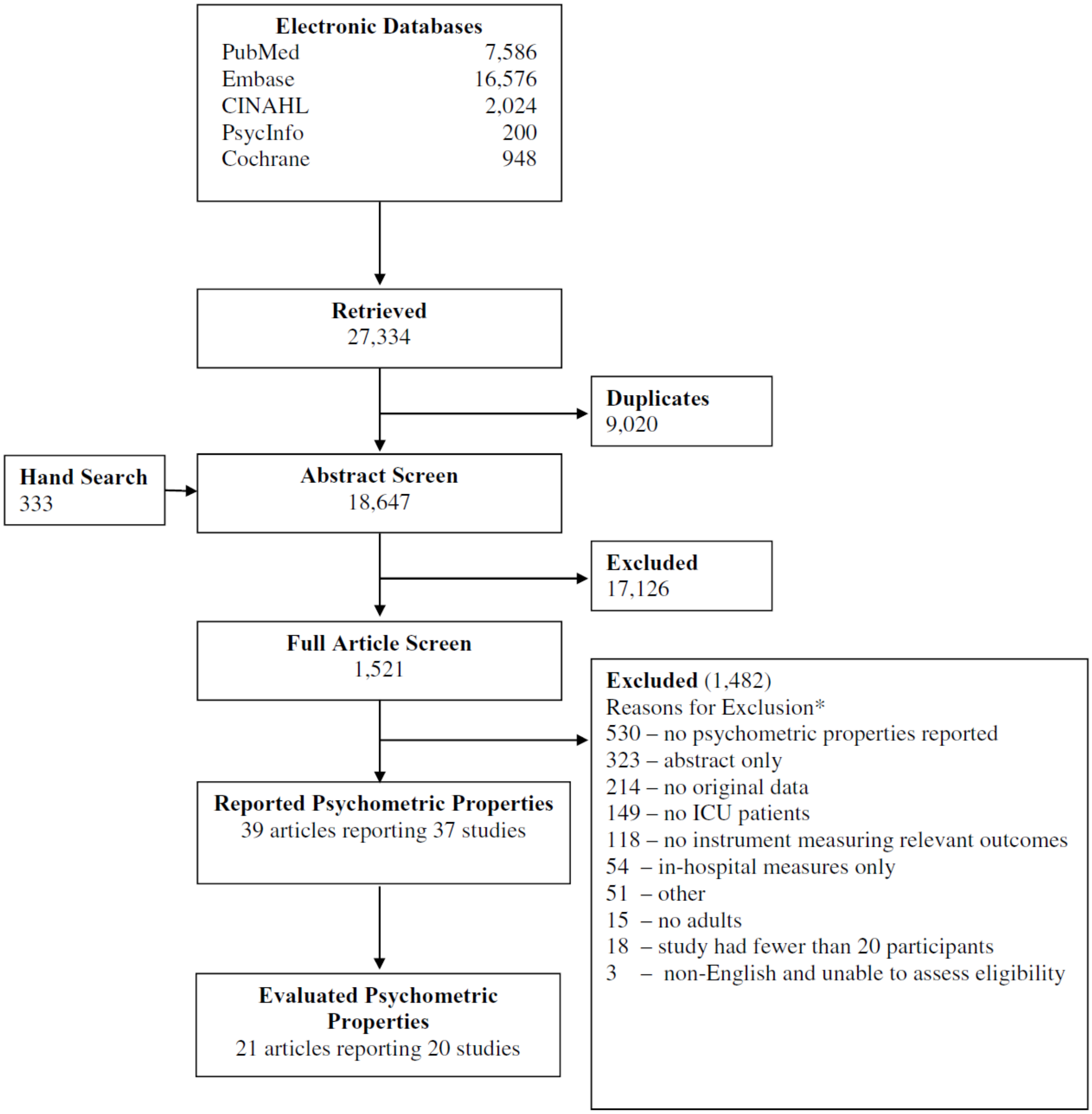
Electronic Databases
PubMed 7,586
Embase 16,576
CINAHL 2,024
PsycInfo 200
Cochrane 948
Retrieved
27,334
Duplicates
9,020
Hand Search
333
Abstract Screen
18,647
Excluded
17,126
Full Article Screen
1,521
Excluded (1,482)
Reasons for Exclusion*
530 – no psychometric properties reported
323 – abstract only
214 – no original data
149 – no ICU patients
118 – no instrument measuring relevant outcomes
54 – in-hospital measures only
51 – other
15 – no adults
18 – study had fewer than 20 participants
3 – non-English and unable to assess eligibility
Reported Psychometric Properties
39 articles reporting 37 studies
Evaluated Psychometric Properties
21 articles reporting 20 studies

Figure 1.

Search flow diagram

* Citations could be excluded for >1 reason

Table 1

Definitions of Measurement Properties of Instruments Assessed by COSMIN Checklist

Measurement Property	Definition
Internal Consistency	The degree of the interrelatedness among the items; the extent to which scores for patients who have not changed are the same using different sets of items from same instrument
Reliability	The proportion of the total variance in the measurements which is due to ‘true’ differences between patients
Measurement Error	The systematic and random error of a patient’s score that is not attributed to true changes in the construct to be measured
Content Validity	The degree to which an instrument measures the construct(s) it purports to measure; the degree to which the content of an instrument is an adequate reflection of the construct to be measured
Structural Validity	The degree to which the scores of an instrument are an adequate reflection of the dimensionality of the construct to be measured
Hypothesis Testing	The degree to which the scores of an instrument are consistent with hypotheses (for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups) based on the assumption that the instrument validly measures the construct to be measured; Idem construct validity
Cross-cultural	Validity The degree to which the performance of the items on a translated or culturally adapted instrument are an adequate reflection of the performance of the items of the original version of the instrument
Criterion Validity	The degree to which the scores of an instrument are an adequate reflection of a ‘gold standard’
Responsiveness	The ability of an instrument to detect change over time in the construct to be measured; Idem responsiveness

Adapted from the COSMIN Checklist Manual (v9)[1]

[1] Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. Journal of Clinical Epidemiology. 2010 Jul;63(7):737-45. PMID: 20494804.

Table 2

Instruments Evaluated by Core Domains

Domain	Type of Measure	Instruments Evaluated
ICF: Body Functions and Structures	Anxiety and depression	DASS; HADS
	Post-traumatic stress disorder	Assessment of traumatic memories; CAPS; IES-R; UK-PTSS14
	Both physical and mental impairment	3Set-4P
ICF: Activity Limitation	Cognitive limitations	Telephone battery
	Physical limitations	6-Minute Walk test
ICF: Participation Restriction	Participation in usual environment	FHS
Quality of Life	Generic quality of life	15D HRQOL; AQOL; EQ-5D; QOL-SP/IT; MSF-36; Satisfaction Scale; SF-36; SF-6D; SIP; Standardized Telephone

ICF: International Classification of Functioning, Disability and Health

15D HRQOL, 15D Health-Related Quality of Life; 3-Set 4P, 3-SET 4P Questionnaire; 6MWT, 6-Minute Walk; AQoL, Assessment of Quality of Life; Assessment of traumatic memories, Assessment of traumatic memories from intensive care unit; CAPS, Clinician-Administered PTSD Scale; DASS, Depression, Anxiety and Stress Scale; EQ-5D, EuroQol - 5 Dimensions; FHS, Follow-up Health Status Questionnaire; HADS, Hospital Anxiety and Depression Scales; IES-R, Impact of Events Scale – Revised; MSF-36, Modified Medical Outcomes Short Form-36; QOL-IT, Italian Quality of Life Questionnaire; QOL-SP, Spanish Quality of Life Questionnaire; Satisfaction Scale, Satisfaction Scale; SF-36 v1, Medical Outcomes Short Form-36 version 1; SF-6D, Medical Outcomes Short Form-6D; SIP,Sickness Impact Profile; Standardized Telephone, Standardized Telephone Interview; Telephone battery, Cognitive Telephone Battery; UK- PTSS-14,UK- Post-Traumatic Stress Syndrome 14-Questions Inventory.

Table 3

Summary of COSMIN Quality Assessment Ratings, by Measurement Property

COSMIN Rating	Internal Consistency	Reliability	Measurement Error	Content validity	Structural Validity	Hypothesis Testing	Cross-Cultural Validity	Criterion Validity	Responsiveness	Total Number of Evaluations (%)
Excellent	0	0	0	0	0	0	0	7	0	7 (4)
Good	0	2	0	0	0	0	0	0	0	2 (1)
Fair	7	34	0	0	10	44	0	10	1	106 (55)
Poor	51	14	0	1	3	6	0	1	3	79 (41)
Total Number (%) of Evaluations	58 (30)	50 (26)	0 (0)	1 (0.5)	13 (7)	50 (26)	0 (0)	18 (9)	4 (2)	194 (100)

Table 4

Characteristics of Studies of Measurement Properties of Instruments Assessing Outcomes in Adult ICU Survivors

Author, Year	Sample Size	Maximum Length of Follow-up (months)	Males, n (%)	Age(Mean (SD))	Population	Illness Severity Measure and Score	Instrument Name	COSMIN Measurement Properties Evaluated
Akerman, 2009	49	12	24(49%)	62(14)	General ICU	APACHE II 18 mean	3-Set 4P	Hypothesis testing, Internal consistency, Reliability, Structural Validity
Akerman, 2013	421	2	254(60%)	68 (15)	General ICU	SAPS 56 mean	3-Set 4P	Hypothesis testing Internal consistency, Reliability, Structural Validity
Alison, 2012	179	6	108†(60%)	57 (16)	NA	APACHE II 19 mean	6MWT	Hypothesis testing.
Bienvenu, 2013	60	60	23(38%)	51 (12)	ARDS	NA	CAP IES-R	Internal Consistency Internal Consistency and Criterion Validity
Capuzzo, 2000	172	12	118(69%)	69 (11)	General ICU	APACHE II 13 mean	QOL-IT QOL-SP	Internal Consistency,Reliability, Hypothesis testing and Responsiveness Internal Consistency, Reliability, Hypothesis testing and Responsiveness
Christie, 2006	79	28	12 (15%)	43 (13)	ARDS	NA	Telephone Battery	Content Validity, Hypothesis testing
Christie, 2006	34	24	17 (50%)	49 (15)	ARDS	APACHE II 19 mean	Telephone Battery	Reliability
Denehy, 2014	177	3	113†(64%)	60 (NR)	ARDS	APACHE II 19 median	6MWT	Hypothesis testing.
Fernandez, 1996	578	6	400(69%)	58 (17)	NA	APACHE II 16 mean	QOL-SP	Internal consistency, Reliability, Structural validity, Hypothesis testing, Responsiveness
Frick, 2002	85	6	NA	65 (NR)	General ICU	SAPS 22 median	Satisfaction Scale SIP Standardized Telephone Interview	Criterion validity, Hypothesis testing, Criterion validity
Heyland, 2000	30	17	16 (53%)†	62 (14)	Sepsis	APACHE II 22 mean	SF-36 v1	Internal consistency, Reliability, Hypothesis testing
Jones, 1993	85	12	49†(58%)	31 (NR)	General ICU	APACHE II 12 median	FHS	Hypothesis testing

Author, Year	Sample Size	Maximum Length of Follow-up (months)	Males, n (%)	Age(Mean (SD))	Population	Illness Severity Measure and Score	Instrument Name	COSMIN Measurement Properties Evaluated
Kaarlola, 2004	2709	72	1788[†] (66%)	54 (NR)	General ICU	APACHE II 14 median	EQ-5D	Hypothesis testing
Khoudri, 2007	179	3	79(54%)	38 (17)	General ICU	APACHE II 14 mean	SF-36 v1	Internal consistency, Reliability
Khoudri, 2012	179	3	79(54%)	38 (17)	General ICU	APACHE II 14 mean	EQ-5D	Reliability, Criterion validity
Lipsett, 2000	128	12	95[†](74%)	58 (17)	General ICU	APACHE II 20 mean	MSF-36 SIP	Internal consistency, Reliability, Structural validity, Hypothesis testing Internal consistency, Reliability, Structural validity
Skinner, 2013	67	6	40[†](60%)	60 (15)	General ICU	APACHE II 17 median	AQoL SF-6D	Internal consistency, Hypothesis testing Internal consistency, Hypothesis testing
Stoll, 1999	52	13 years[*]	26(50%)	36 (18-50)[‡]	ARDS	APACHE II 22 median	Assessment of Traumatic Memories	Internal consistency, Reliability, Criterion validity
Sukantarat, 2007	51	9	22(43%)	57 (14)	General ICU	APACHE II 14 median	DASS HADS	Internal validity,Criterion validity Internal validity
Twigg, 2008	56	3	20(45%)	56 (NR)	General ICU	APACHE II 16 median	UK-PTSS_14	Internal consistency, Reliability, Hypothesis testing, Criterion validity
Vainiola, 2010	3600	12	NA	NA	General ICU	NA	15D HRQOL	Hypothesis testing, Responsiveness

NA = Not Available, NR = Not Reported

APACHE, Acute Physiology and Chronic Health Evaluation; SAPS, Simplifies Acute Physiology Score
ARDS, acute respiratory distress syndrome

15D HRQOL, 15D Health-Related Quality of Life; 3-Set 4P, 3-SET 4P Questionnaire; 6MWT, 6-Minute Walk; AQoL, Assessment of Quality of Life; Assessment of traumatic memories, Assessment of traumatic memories from intensive care unit; CAPS, Clinician-Administered PTSD Scale; DASS, Depression, Anxiety and Stress Scale; EQ-5D, EuroQol - 5 Dimensions; FHS, Follow-up Health Status Questionnaire; HADS, Hospital Anxiety and Depression Scales; IES-R, Impact of Events Scale – Revised; MSF-36, Modified Medical Outcomes Short Form-36; QOL-IT, Italian Quality of Life Questionnaire; QOL-SP, Spanish Quality of Life Questionnaire; Satisfaction Scale, Satisfaction Scale; SF-36 v1, Medical Outcomes Short Form-36 version 1; SF-6D, Medical Outcomes Short Form-6D; SIP,Sickness Impact Profile; Standardized Telephone, Standardized Telephone Interview; Telephone battery, Cognitive Telephone Battery; UK-PTSS-14, UK- Post-Traumatic Stress Syndrome 14-Questions Inventory.

[*] 1st interview median 4 years from ICU discharge (first = 0 – last = 10 years) with median difference of 2 years to the second interview

[†] Values calculated based on provided data

‡Median age and numbers in parentheses are the minimum and maximum age