Bhaloo Z, Adams D, Liu Y, Hansraj N, Hartling L, Terwee CB, Vohra S

Title: Primary Outcomes Reporting in Trials (PORTal): a systematic review of inadequate reporting in pediatric randomized controlled trials

AUTHOR POST PRINT VERSION

Bhaloo Z, Adams D, Liu Y, Hansraj N, Hartling L, Terwee CB, Vohra S. Primary Outcomes Reporting in Trials (PORTal): a systematic review of inadequate reporting in pediatric randomized controlled trials. *Journal of Clinical Epidemiology*. 2017 Jan 31;81:33-41.

Keywords: pediatric; outcome measure; primary outcome; reporting; systematic review; randomized controlled trial

Abstract

Objective: Conduct a systematic review of pediatric randomized controlled trials published in high impact journals to assess the reporting of primary outcomes and the psychometric properties of their measures.

Study Design and Setting: Systematic review with screening and simultaneous data extraction conducted by two independent reviewers. Electronic searches of six general medicine and four pediatric journals were conducted in MEDLINE, EMBASE and the Cochrane Central Register of Controlled Trials (CENTRAL) databases. Randomized controlled trials of a single phase/step in a single publication, published in English between 2000 and 2010 with participants less than 21 years of age were included.

Results: A random sample of 20% (n=446) of 2229 initial references was screened and 206 (46%) met inclusion criteria. Half (48.5%) of included studies reported a singular primary outcome, 27% did not identify any primary outcome, and 24% identified multiple primary outcomes (range 2-20). Twenty one trials used an instrument to measure their primary outcome, but only 7 (33%) reported its psychometric properties.

Conclusions: Pediatric trials published in top medical journals have inadequate reporting of their primary outcomes and the psychometric properties of their outcome measures. Whether the issue is one of poor reporting and/or poor validation will be further investigated.

Keywords: pediatric; outcome measure; primary outcome; reporting; systematic review; randomized controlled trial

Title: Primary Outcomes Reporting in Trials (PORTal): a systematic review of pediatric randomized controlled trials

ADDS TO WHAT WAS KNOWN

Despite widespread acceptance of reporting guidelines for clinical trials, preliminary investigation in select populations has identified that primary outcomes are poorly reported. Reporting guidelines represent a "minimum set" and do not address measurement properties of primary outcome measures. Since trials are only as valid as their primary outcome measures, adequate reporting of measurement properties is essential.

KEY FINDING

This review examines a cross-section of pediatric trials published in high impact journals. It reveals that pediatric trials, across disciplines, have inadequate reporting of both their primary outcomes and the measurement properties of their outcome measures.

WHAT IS THE IMPLICATION, WHAT SHOULD CHANGE NOW

Trialists, journal editors, and reporting guideline developers should work together to improve reporting of primary outcome measures and their measurement properties

BACKGROUND

Randomized controlled trials (RCTs) represent the gold standard for evidence about treatment effectiveness for health care providers, researchers, policy-makers and other decision-makers. RCTs are preferentially included in knowledge synthesis efforts such as systematic reviews and meta-analyses, which inform decision-makers at every level. Many RCTs are published annually in high impact journals; however, there is growing concern with regards to the reporting of outcomes and consequently the reporting of the measurement properties of the outcome measures, namely their validity and reliability.[1-4] As clinical trials are "only as credible as their outcomes"[5], a lack of reporting and validation implies that tremendous expense, effort, and resources may not be used optimally.

An outcome is a measurable variable that should be clearly stated by the authors and an outcome measure is the tool used for measuring the outcome (scales, questionnaires, instruments, or scoring systems – we describe these collectively using the term "outcome measure").[1] The measurement properties of an outcome measure, i.e. validity, reliability and responsiveness provide information regarding the measure's intended purpose, its performance and accuracy, and its ability to detect a true change. When selecting which outcome measures to use in any given study or when evaluating the use of a particular measure, the measurement properties are often compared. Inadequacies related to primary outcome reporting and their consequent impediment on the conduct of knowledge synthesis efforts has been discussed in light of selective outcomes reporting, wherein only a selected subset of analyses or outcomes are reported based on the results they yield.[6]

The issue of selective outcomes reporting is secondary to a larger issue of trials that fail to identify any primary outcome at all. The inadequate reporting of outcomes in the pediatric population has been identified while investigating outcomes selection within a specified clinical area. In systematic reviews of RCTs within pediatric subspecialties, authors consistently fail to report identifiable primary outcomes.[1,4,7]

Although it is recognized that the "prespecification of a single primary outcome based on biologic credibility, clinical importance, and potential responsiveness to the intervention" is the best approach, the reader is more often "offered a shopping list of end points"[4]. Along with the poor reporting of primary outcomes, the validation of outcome measures is also poorly reported or missing altogether. Few studies report that a validated instrument was used or provide evidence of formal evaluation against some sort of reference standard, and those that do, fail to provide citations to support the reported measurement properties.[1,3]

A variety of initiatives[5, 8-10] have been developed along with systematic reviews[6, 7] that address some of the issues of inadequate reporting and validation. To assess the magnitude of this problem across pediatric disciplines, we conducted a systematic review of a random sample of pediatric RCTs published in ten high impact journals between 2000 and 2010. Our primary interest was assessing outcome measures, since these have been identified as in need of further study. As such, the main aim or primary outcome of this systematic review was to examine primary outcome reporting including: (1) how many RCTs reported a primary outcome, (2) the number of primary outcomes reported, (3) how many RCTs reported the measurement properties of the instruments used, and (4) the relevant citations provided for the measurement properties reported. A secondary outcome was to examine other key pediatric trial metrics and their reporting, such as information about the population (participant ages, condition(s) under study, sample size and calculation), intervention and control group(s).

METHODS

Search Strategy

With the help of an experienced health research librarian, electronic searches in MEDLINE, EMBASE and the Cochrane Central Register of Controlled Trials (CENTRAL) databases were undertaken. We selected 10 journals by impact factor (six general medicine journals and four pediatric journals), all of which include pediatric trials in their publications. All searches used the respective journals name: *New England Journal of Medicine, Journal of the American Medical Association, Lancet, Annals of Internal Medicine, British Medical Journal, Plos Medicine,*

Journal of the American Academy of Child and Adolescent Psychiatry, *Pediatrics*, *Journal of Pediatrics*, and *Archives of Pediatrics & Adolescent Medicine*. Searches were limited by publication type (RCTs), publication year (2000-2010), respective pediatric filters, and the English language. The full search strategies for each database are available by request to the corresponding author.

Study Selection

We included studies that (1) were RCTs, i.e. studies that randomly allocated participants to interventions, and included parallel, cross-over, factorial or N-of-1 designs, (2) were comprised of a single phase (or single step intervention) in a single publication as it is difficult to extract data for trials with multiple phases and steps that may contain different methods/interventions/outcomes in each phase/step (multiple steps may also result in multiple primary outcomes and thereby skew our findings), (3) included only a pediatric population (less than 21 years of age) as it is unlikely outcome measures have been validated for both adults and children, (4) were of any intervention type, and (5) were published in one of the previously identified ten high impact journals between 2000 and 2010. We excluded: (1) studies that were diagnostic or screening in nature as this initiative was focused on improving reporting based on CONSORT guidelines, and other reporting guidelines exist for diagnostic and screening studies; (2) self-described pilot studies, which may not place the same emphasis on primary outcome measure selection and reporting, and (3) follow up studies or secondary publications . An a priori decision was made to select a random sample (20%) of studies. The articles were listed in a Microsoft Excel spreadsheet and randomly ordered using the RAND and SORT functions and 20% were selected. The articles were therefore not stratified by journal. The titles and abstracts of the 20% sample were then screened by independent reviewers (ZB, YL, NH) for potential inclusion. Full texts of the selected articles were then retrieved and each article was independently assessed by the same reviewers for inclusion based on the pre-defined criteria. Disagreements were resolved with a senior team member (DA or SV) and consensus was reached.

Data Extraction

Screening and data extractions were carried out by three independent reviewers (ZB, YL, NH) using a standardized data extraction form that was piloted prior to use. Extracted variables were entered into Distiller SR (web-based systematic review software) and included: journal, publication year, design of RCT, population age, condition and intervention of interest, sample size and sample size calculation, number of primary outcomes, outcome measures used and details of their measurement properties.

An explicit report or reference to a primary outcome was searched for in the abstract and full text of all included studies. As an additional measure, the "find" tool was also used to identify any mention of a primary outcome or similar terminology within the text. As per the CONSORT statement,[11] "the primary outcome measure is the pre-specified outcome considered to be of greatest importance to relevant stakeholders". Great flexibility of terminology was allowed for in the identification of primary outcome(s) (e.g. main outcome, primary end point, primary objective) and specific terminology used was recorded.

Studies reporting a single primary outcome were further assessed for the report of an outcome measure. An outcome measure is identified as "a scale, scoring system, instrument, questionnaire or other tool used for measuring an outcome. [1]" Measurement properties of the outcome measures reported were identified based on the COSMIN group's (COnsensus-based Standards for the selection of health Measurement INstruments) published standardized terminology, definitions, and taxonomy of measurement properties for the evaluation of instruments based on international consensus.[9] In addition to reading the text, the "find" tool was used to identify any mention of a measurement property within the text. For studies reporting on the measurement properties of the outcome measures used, citations and bibliographies were searched for evidence to support these reports and the age groups for whom the properties were studied were also compared to the age groups in which the measures were used. Any discrepancies in data extraction were noted and resolved through joint discussions with a senior team member (DA or SV).

Data Analysis

Systematic review methods and reporting were completed according to PRISMA guidelines.[12] This systematic review does not evaluate the effectiveness or safety of a particular intervention but rather focuses on reporting, therefore risk of bias and meta-analysis are not necessary or relevant. Data were entered into DistillerSR and analyzed using STATA. Results are described using descriptive statistics (summary scores, proportion, frequency) and presented as percentages.

RESULTS

Our electronic search yielded 2229 unique references (Figure 1-1). The titles and abstracts of a random 20% sample (n=445) were screened; of these 445, 173 were excluded. Of these 173 excluded, 29 were not RCTs, 44 articles were follow-up studies, 10 were pilot studies, 12 articles reported on more than one phase/step/trial, two were diagnostic and screening trials, 70 studies also included adults, four articles were not retrievable, and two were duplicate articles. The full text of 272 potentially relevant studies was retrieved and screened. Of these 272 articles, 66 were excluded: 3 were not RCTs; 6 were follow-up studies; 2 were pilot studies; 12 reported on more than one phase/step/trial; 42 also included adults; and one provided no age information for participants. A total of 206 RCTs were included for data extraction.

Of the included studies, 77 % were from pediatric journals; 32% were from *Pediatrics*, 28% from *the Journal of Pediatrics*, 10% from the *Archives of Pediatrics and Adolescent Medicine*, and 7% from the *American Academy of Child & Adolescent Psychiatry* while the remaining 23% were published in the *Lancet* (9%), the *New England Journal of Medicine* (9%) the *Journal of the American Medical Association* (4%), *PLoS Medicine* (1%), the *British Medical Journal* (0.5%), and the *Annals of Internal Medicine* (0.5%). Of the 206 RCTs, 89% were parallel in design and the remainder were crossover or factorial trials. The majority (65%) were treatment trials as opposed to prevention trials (35%). A median of two groups were studied in each trial (range 2-6). A variety of conditions were studied across the 206 trials and these included: type 1 diabetes, respiratory distress syndrome, patent ductus arteriosus, obesity, Kawasaki disease, bronchiolitis, cystic fibrosis, depression, asthma, and bronchopulmonary dysplasia. Only 63% of

RCTs provided a sample size calculation and sample sizes ranged from 10 to 63 225 participants (median = 120, IQR = 321). Most authors did not explicitly report actual age ranges (upper and lower bounds) of their participants but rather provided the mean age of their population. Variables extracted from the included studies are summarized in Table 1-1.

Primary Outcomes

A variety of terminology for "primary outcome" were used, including primary outcome(s), primary endpoint(s), primary efficacy variable(s), main outcome measure(s), primary study variable(s), primary outcome measure(s), primary study end point(s), primary outcome variable(s), primary objective(s), primary pre-specified outcome(s), primary dependent variable(s), main outcome measurement(s), and primary efficacy parameter(s).

Of the 206 RCTs included, 100 (48.5%) explicitly reported a single primary outcome, 56 (27.2%) did not identify any primary outcome, and 50 studies (24.3%) identified multiple primary outcomes. The 50 studies that reported multiple primary outcomes identified two to 20 outcomes as primary with a median of two primary outcomes (IQR 1).

Outcome Measures

Of the 100 studies that reported a single primary outcome, 19 reported the use of an outcome scale, tool, or instrument to measure their primary outcome (Table 1-2). The other 81 studies used physiologic measures (eg. eosinophil-derived neurotoxin levels, calcium absorption, rate of decline in forced expiratory volume), diagnostic tools (eg. polysomnography, radiology), or quantitative indexes such as duration of stay in hospital to measure their primary outcome and were thus not evaluated further as regulations and accreditations exist for such measures.

Of the 19 studies reporting the use of an outcome measure, seven (37%) reported measurement properties. All seven studies provided relevant citations to support their reports and three (43%) examined the measurement properties as part of their study. For the 12 studies that did not explicitly report any measurement properties, any citations provided for the outcome measures themselves were reviewed. We found that the outcome measure citations provided in 11 (92%) of the 12 studies were in fact relevant citations for measurement properties.

DISCUSSION

More than 10 years after CONSORT, one quarter (27.2%) of pediatric RCTs published in high impact journals fail to report any primary outcome. This is especially surprising as all of the journals included in our review have endorsed CONSORT. Furthermore, measurement properties of outcome measures are often not reported by authors although these measures are used to evaluate the trial's primary outcome. Although relevant citations evaluating the measurement properties of these outcome measures were available for the majority of instruments (92%), they are often not reported by authors (63%). Since RCTs are "only as credible as their outcomes"[5], it is crucial that their outcomes are valid and reliable in the population in which they are being applied, and clearly reported as such.

The results of this study may be limited in part due to the methods used to search for our included studies. We recognize that assessment of reporting in only high impact journals may lead to an underestimate of the problem, however we chose this as our sample since knowledge users are more likely to be convinced of our findings if they cannot discount them due to their lack of familiarity with smaller journals (i.e. journals that they do not hold in high regard or aspire to publish in). High impact journals are presumed to have the most rigorous and stringent publication standards so if a significant problem exists in this group then our findings likely under-estimate the extent of the problem in lower impact journals and grey literature. Furthermore, as we did not know the extent of the problem of reporting, our ability to perform a sample size calculation was limited. We chose to assess a random 20% sample as we believe this represents a comprehensive and feasible sample of pediatric RCTs across disciplines and journals.

Strengths of our approach include use of systematic review methods and reporting according to PRISMA guidelines.[12] Reviewers independently screened and extracted data from the studies using standardized forms. This systematic review also accepted a wide range of terminology for the reporting of a primary outcome. By recognizing the variety of terms used to identify a primary outcome, we avoid over-estimation and provide a clearer, fairer picture of the scope of the problem. Great heterogeneity exists in the author descriptions of primary outcomes. Of note, authors use "outcome" and "outcome measure" interchangeably. It is suggested that an outcome

is a measurable variable while the outcome measure is the tool used for measuring the outcome (such as scales, questionnaires, instruments, or scoring systems).[1] The inconsistency and heterogeneity of these terms across initiatives and organizations does not aid in clarity and we suggest that it is time for trialists, editors, and guideline developers to reach consensus on acceptable terminology. Regardless of terminology, primary outcomes are not explicitly being reported in RCT publications.

Although other studies have explored the issues of outcomes reporting and the validation of instruments, this has been limited in scope and restricted to individual disciplines. [1-4] A thorough synthesis of the problem across disciplines had not previously been conducted. To our knowledge, our systematic review of pediatric RCTs in high impact journals is the first of its kind to look specifically at the problem of reporting and validation of primary outcomes and their measures.

RCTs are heavily relied upon by evidence-based decision makers, researchers, funding agencies, policy makers, peer reviewers, authors and journal editors. A substantial proportion of RCTs fail to report a single primary outcome and too often, measurement properties of measures are also unreported. The validity of these trials is directly reflected by the validity of the primary outcomes and the measures used. The results of this review can be used to improve reporting standards by facilitating the revision of reporting guidelines such that they require the clear reporting of a study's primary outcome and relevant citations for measurement properties of outcome measures. Given the extent of the problem, we recommend that future iterations of trial reporting guidelines (i.e., CONSORT) specify that trials must report a primary outcome and its measurement properties. This study may also aid in the informed selection of outcomes and outcome measures by trialists and other clinical researchers. The research findings presented here have the potential to encourage higher standards for the reporting and conduct of trials such that RCT results can be used more confidently at every level of knowledge synthesis and translation.

While we have firmly identified the inadequate reporting of primary outcomes in pediatric RCTs, the issues around reporting of measurement properties need further investigation. Lack of reporting of these properties may reflect the failure of authors to explicitly report known measurement properties or there may be a lack of formal assessment of these outcome measures

therefore limiting the ability to report measurement properties. Whether the issue is one of inadequate reporting, insufficient validation or a combination of the two needs to be determined.

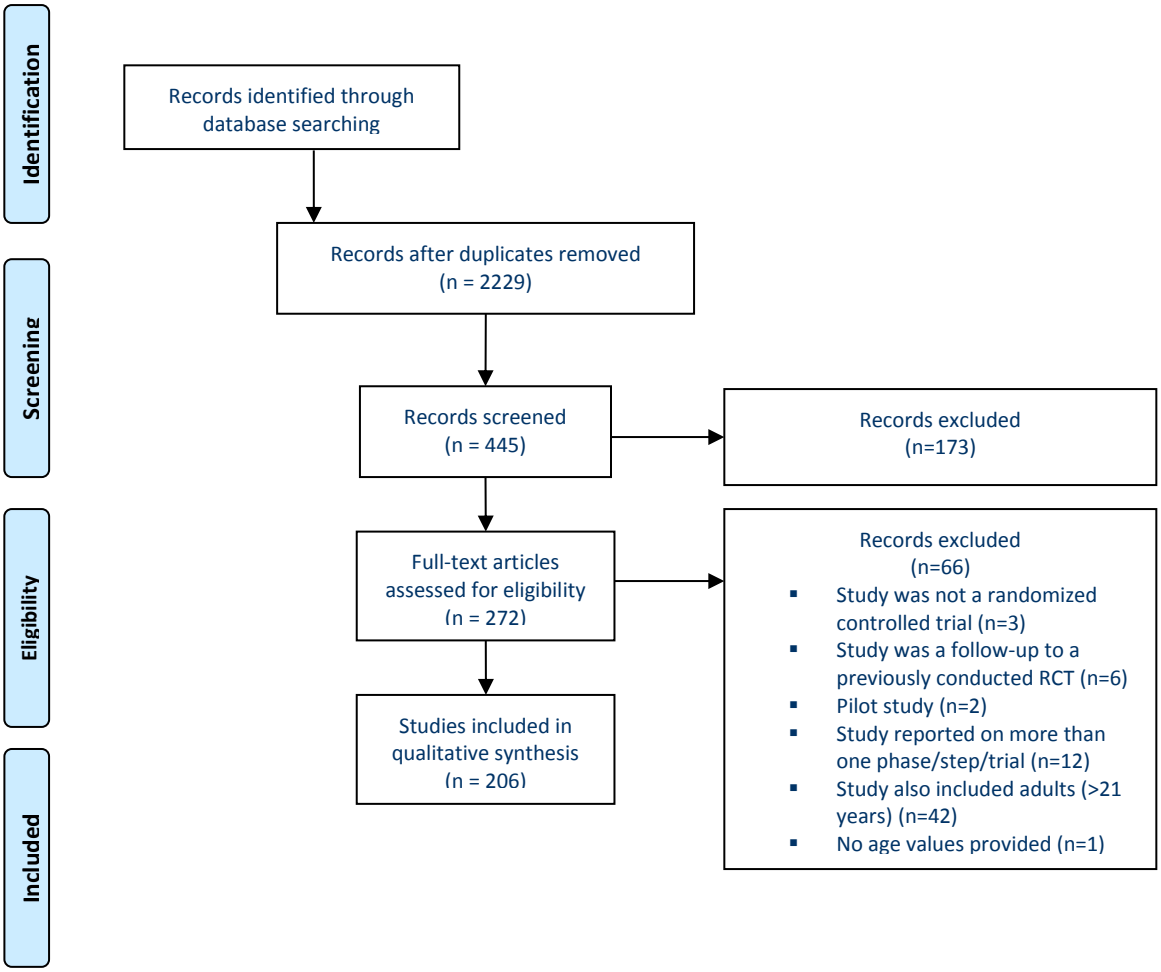

Figure 1-1 PRISMA[12] Flow Diagram of Search Results

Variable		Number of RCTs (n=206)
Publication Year	2000	18 (9%)
	2001	18 (9%)
	2002	16 (8%)
	2003	27 (13%)
	2004	19 (9%)
	2005	26 (13%)
	2006	23 (11%)
	2007	20 (10%)
	2008	17 (8%)
	2009	18 (9%)
	2010	4 (2%)
Journal	*Pediatrics*	65 (32%)
	Journal of Pediatrics	57 (28%)
	Archives of Pediatrics and Adolescent Medicine	20 (10%)
	Lancet	18 (9%)
	New England Journal of Medicine	18 (9%)
	American Academy of Child & Adolescent Psychiatry	15 (7%)
	Journal of the American Medical Association	9 (4%)
	PLoS Medicine	2 (1%)
	British Medical Journal	1 (0.5%)
	Annals of Internal Medicine	1 (0.5%)
Type of RCT	Parallel	183 (89%)
	Crossover	20 (10%)
	Factorial	3 (1%)
	N-of-1	0 (0%)
Type of trial	Treatment	134 (65%)
	Prevention	72 (35%)
Number of groups studied	Median	2
	Range	2-6
Sample Size Calculation	Reported	131 (64%)
Sample Size	Median	120
	Range	10 – 63 225

Table 1-1 Summary of Included Studies

Primary Outcome	Outcome Measure	Measurement Properties reported	Authors' Citations for Measurement Properties
Changes in the retractions and wheezing in acute bronchiolitis	Respiratory Disease Assessment Instrument (RDAI) – Respiratory Assessment Change Score[13]	Internal validity and responsiveness	Klassen T, Sutcliffe T, Watters L, Wells GA, Allen UD, Li MM. Dexamethasone in albuterol-treated inpatients with acute bronchiolitis: a randomized, controlled trial. J Pediatr 1997;130:191-7. Klassen TP, Rowe PC, Sutcliffe T, Ropp LJ, McDowell IW, Li MM. Randomized trial of albuterol in acute bronchiolitis. J Pediatr 1991;118:806-11. Lowell DI, Lister G, Von Koss H, Mc-Carthy P. Wheezing in infants: the response to epinephrine. Pediatrics 1987; 79:939-45.
Proportion of treatment successes (i.e. need for enteral feeding in infants with resistance to feeding)	Infant Feeding Behaviours - Rater checklist (IFB – Rater checklist)[14]	Previously validated, √agreement between raters	Arts-Rodas D, Benoit D. Feeding problems in infancy and early childhood: identification and management. J Paediatr Child Health 1998; 3:21-7. Benoit D, Green D. The Infant Feeding Behaviors - Rater Checklist: preliminary data. Poster presented at the Fortysecond Annual Meeting of the American Academy of Child and Adolescent Psychiatry, New Orleans, LA; 1995. Koulis K, Arts-Rodas D, Benoit D. The Infant Feeding Behaviors - Rater checklist: comparison of coding methods. Poster presented at the forty-fourth Annual Meeting of the American Academy of Child and Adolescent Psychiatry,Toronto, Ontario; 1997.
Adequate clinical response defined by depressive symptoms	Children's Depression Rating Scale - Revised (CDRS-R)[15]	√inter-rater reliability, intra-class correlation	Poznanski EO, Freeman LN, Mokros HB. Children's Depression Rating ScaleYRevised. Psychopharmacol Bull. 1984;21:979Y989. Guy W. ECDEU Assessment Manual for Psychopharmacology. 2nd ed. Washington: US Government Printing Office; 1976.
Adequate clinical response defined by depressive symptoms Exacerbation rates in lithium treatment of acute mania	Clinical Global Impressions-Improvement Subscale (CGI-I)[15, 17]	√inter-rater reliability, intra-class correlation -	Poznanski EO, Freeman LN, Mokros HB. Children's Depression Rating ScaleYRevised. Psychopharmacol Bull. 1984;21:979Y989. Guy W. ECDEU Assessment Manual for Psychopharmacology. 2nded. Washington: US Government Printing Office; 1976. -
ADHD Symptoms	Attention-deficit/Hyperactivity Disorder Rating Scale-IV-Teacher Version: Investigator administered and scored (ADHDRS-IV-Teacher:Inv)[16]	validity	Faries DE, Yalcin I, Harder D, Heiligenstein JH (2001), Validation of the ADHD Rating Scale as a clinician administered and scored instrument. J Atten Disord 5:39–47.

Exacerbation rates in lithium treatment of acute mania	Global Clinical Judgements (GCJ)[17]	-	(Campbell M, Small AM, Green WH et al. (1984), Behavioral efficacy of haloperidol and lithium carbonate: a comparison in hospitalized aggressive children with conduct disorder. *Arch Gen Psychiatry* 41:650–656. Campbell M, Adams P, Small AM et al. (1995), Lithium in hospitalized aggressive children with conduct disorder: a double-blind and placebocontrolled study. *J Am Acad Child Adolesc Psychiatry* 34:445–453. Malone RP, Delaney MA, Luebbert JF, Cater J, Campbell M (2000), A double-blind placebo-controlled study of lithium in hospitalized aggressive children and adolescents with conduct disorder. *Arch Gen Psychiatry* 57:649–654)
Pain induced by heel lance in newborns	Premature Infant Pain Profile (PIPP)[18]	validated, interrater reliability	Ballantyne M, Stevens B, McAllister M, Dionne K, Jack A. Validation of the premature infant pain profile in the clinical setting. *Clin J Pain.* 1999;15(4):297–303. Jonsdottir RB, Kristjansdottir G. The sensitivity of the premature infant pain profile: PIPP to measure pain in hospitalized neonates. *J Eval Clin Pract.* 2005;11(6):598–605.
Duration of acute viral upper respiratory tract infection	Canadian Acute Respiratory Illness and Flu Scale (CARIFS)[19]	-	(Jacobs B, Young NL, Dick PY, et al. Canadian Acute Respiratory Illness and Flu Scale (CARIFS): development of a valid measure for childhood respiratory infections. *J Clin Epidemiol* 2000; **53:**793–99.)
Gross motor function	Gross Motor Function Measure (GMFM)[20]	-	(Russell DJ, Rosenbaum PL, Cadman DT, Gowland C, Hardy S, Jarvis S. The Gross Motor Function Measure: a means to evaluate the effects of physical therapy. *Develop Med Child Neurol* 1989; **31:** 341–52. Nordmark E, Hagglund G, Jarnlo GB. Reliability of the gross motor function measure in cerebral palsy. *Scand J Rehab Med* 1997; **29:** 25–28. Trahan J, Malouin F. Changes in gross motor function measure in children with different types of cerebral palsy: an eight month follow-up study. *Pediatr Phys Ther* 1999; **11:** 12–17.)
Composite of death or severe neurodevelopmental disability[21] Composite of death, cerebral palsy, cognitive delay, deafness, or blindness[22]	Gross Motor Function Classification System (GMFCS) [21,22]		(Palisano RJ, Hanna SE, Rosenbaum PL, et al. Validation of a model of gross motor function for children with cerebral palsy. Phys Ther 2000;80:974-85.) (Palisano R, Rosenbaum P, Walter S, Russell D, Wood E, Galuppi B. Development and reliability of a system to classify gross motor function in children with cerebral palsy. Dev Med Child Neurol 1997; 39:214-23.)
Composite of death or severe neurodevelopmental disability[21]	Mental Developmental Index of the Bayley Scales of Infant Development II (BSID-II)[21, 22]	-	(Bayley N. Bayley scales of infant development. 2nd ed. San Antonio, TX: Psychological Corporation, 1993.) (Bayley N. Manual for the Bayley Scales of Infant Development. 2nd ed. San Antonio, TX: Psychological Corporation, 1993.

Composite of death, cerebral palsy, cognitive delay, deafness, or blindness[22]			Hack M, Taylor G, Drotar D, et al. Poor predictive validity of the Bayley Scales of Infant Development for cognitive function of extremely low birth weight children at school age. Pediatrics 2005;116:333-41.)
Symptoms of obsessive-compulsive disorder (change in score from baseline)[23, 24]	Children's Yale-Brown Obsessive-Compulsive Scale (CY-BOCS)[23, 24]	-	(Scahill L, Riddle MA, McSwiggin-Hardin M et al. (1997), Children's Yale- Brown Obsessive Compulsive Scale: reliability and validity. *J Am Acad Child Adolesc Psychiatry* 36:844–852)
Severe deformational plagiocephaly	Oblique Diameter Difference Index (ODDI)[25]	-	(van Vlimmeren LA, Takken T, van Adrichem LN, van der Graaf Y, Helders PJ, Engelbert RH. Plagiocephalometry: a non-invasive method to quantify asymmetry of the skull; a reliability study. *Eur J Pediatr*. 2006;165(3):149-157.)
Difference in performance on tests assessing cognitive functions in children with Down syndrome	ǂCognitive Test Battery[26] - Stroop Color/Shape - Stroop Color/Word - Auditory Continuous Performance Task (ACPT) - Visual Continuous Performance Task - McCarthy Scales of Children's Abilities - Wide Range Assessment of Memory and Learning (WRAML) - Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI-R) - Delayed match-to-sample - Match-to-sample - Go/No-go - Wide Range Assessment of Visuo-Motor Abilities (WRAVMA)		(Johnson CJ. Effects of color on children's naming of pictures.Percept Mot Skills. 1995;80:1091-1101. Dalton AJ. Dementia in Down syndrome: methods of evaluation. In: Nadel L, Epstein CJ, eds. Down Syndrome and Alzheimer Disease. New York, NY: Wiley- Liss Inc; 1992:51-76. McCarthy D. McCarthy Scales of Children's Abilities. New York, NY: Psychological Corp; 1972. Sheslow D, Adams W. Wide Range Assessment of Memory and Learning. Wilmington, Del: Jastak Associates Inc; 1990. Wechsler D. Manual for the Wechsler Preschool and Primary Scale of Intelligence. San Antonio, Tex: Psychological Corp; 1967. Adams W, Sheslow D. Wide Range Assessment of Visuo-Motor Abilities. Wilmington, Del: Wide Range Inc; 1995.)
Change in individual test scores to assess safety	evaluation tool designed by authors[27]	(authors indicate it has not been	-

knowledge		validated)	
Physical self-worth in obesity	Children and Youth Physical Self-Perception Profile (CY-PSPP)[28]	-	(Whitehead JR. A study of children's physical self-perceptions using an adapted physical self-perception profile questionnaire. *Pediatr Exerc Sci.* 1995;7:132–151. Biddle S, Page A, Ashford B, et al. Assessment of children's physical self-perceptions. *Int J Adolesc Youth.* 1993;4:93–109)
Anxiety of the child	Modified Yale Preoperative Anxiety Scale (m-YPAS)[29]	reliability and validity	Kain ZN, Mayes LC, Cicchetti DV, Bagnall AL, Finley JD, Hofstadter MB. The Yale Preoperative Anxiety Scale: how does it compare with a "gold standard?*" Anesth Analg.* 1997;85:783–788.
Change in irritability from baseline	Aberrant Behaviour Checklist (ABC)[30]	-	(Aman MG, Singh NN, Stewart AW, Field CJ. The aberrant behaviour checklist: a behavior rating scale for the assessment of treatment effects.*Am J Ment Defic.* 1985;89:485–491. Aman MG, Singh NN. *Aberrant Behavior Checklist Manual*. East Aurora, NY: Slosson Educational Publications; 1986)
Neurobehavioral development	Neurobehavioural Assessment of the Preterm Infant (NAPI)[31]	√test-retest reliability, interrater reliability, clinical validity and sensitivity	(Korner AF, Kraemer HC, Reade EP, Forrest T, Dimiceli S, Thom VA. A methodological approach to developing an assessment procedure for testing the neurobehavioral maturity of preterm infants. *Child Dev.* 1987;58:1478–1487) Korner AF, Constantinou J, Dimiceli S, Brown BW, Thom VA. Establishing the reliability and developmental validity of a neurobehavioral assessment for preterm infants: a methodological process. *Child Dev.* 1991;62:1200–1208. Korner AF, Stevenson DK, Kraemer HC, et al. Prediction of the development of low birth weight preterm infants by a new neonatal medical index. *Dev Behav Pediatr.* 1993;14:106–111.

Table 1-2 Outcome Measures and Measurement Properties

ǂ battery test: comprises of 14 tests/domains selected from a variety of measures - treated as one outcome measure
( ) was not referred to by authors in text but included in the bibliographies
√ at least one of the measurement properties was examined as part of the study

References

1. Johnston BC, Shamseer L, da Costa BR, Tsuyuki RT, Vohra S. Measurement issues in trials of pediatric acute diarrheal diseases: a systematic review. Pediatrics 2010;126:e222-31.

2. Sinha I, Jones L, Smyth RL, Williamson PR. A systematic review of studies that aim to determine which outcomes to measure in clinical trials in children. PLoS Med 2008;5:e96.

3. Reid GT, Walter FM, Brisbane JM, Emery JD. Family history questionnaires designed for clinical use: a systematic review. Public Health Genomics 2009;12:73-83.

4. Zhang B, Schmidt B. Do we measure the right end points? A systematic review of primary outcomes in recent neonatal randomized clinical trials. J Pediatr 2001;138:76-80

5. Tugwell P, Boers M. OMERACT conference on outcome measures in rheumatoid arthritis clinical trials: introduction. J Rheumatol 1993;20:528-30.

6. Kirkham JJ, Dwan KM, Altman DG, Gamble C, Dodd S, Smyth R, Williamson PR. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. BMJ 2010;340:c365.

7. Sinha IP, Williamson PR, Smyth RL. Outcomes in clinical trials of inhaled corticosteroids for children with asthma are narrowly focussed on short term disease activity. PLoS One 2009;4:362-76.

8. Moher D, Jones A, Lepage L. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. JAMA 2001;285:1992-5.

9. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010;63:737-45.

10. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34-42.

11. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. Ann Intern Med 2010;152:726-32.

12. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009;6:e1000097.

13. Schuh S, Coates AL, Binnie R, Allin T, Goia C, Corey M, et al. Efficacy of oral dexamethasone in outpatients with acute bronchiolitis. J Pediatr 2002;140:27-32.

14. Benoit D, Wang EL, Zlotkin SH. Discontinuation of enterostomy tube feeding by behavioural treatment in early childhood: a randomized controlled trial. J Pediatr 2000;137:498-503.

15. Asarnow JR, Emslie Gr, Clarke G, Wagner KD, Spirito A, Vitiello B, et al. Treatment of selective serotonin reuptake inhibitor-resistant depression in adolescents: predictors and moderators of treatment response. J Am Acad Child Adolesc Psychiatry 2009;48:330-9.

16. Weiss M, Tannock R, Kratochvil C, Dunn D, Velez-Borras J, Thomason C, et al. A randomized, placebo-controlled study of once-daily atomoxetine in the school setting in children with ADHD. J Am Acad Child Adolesc Psychiatry 2005;44:647-55.

17. Kafantaris V, Coletti DJ, Dicker R, Padula G, Pleak RR, Alvir JMJ, et al. Lithium treatment of acute mania in adolescents: a placebo-controlled discontinuation study. J Am Acad Child Adolesc Psychiatry 2004;43:984-93.

18. Codipietro L, Ceccarelli M, Ponzone A. Breastfeeding or oral sucrose solution in term neonates receiving heel lance: a randomized, controlled trial. Pediatrics 2008;122:e716-21.

19. Butler CC, Robling M, Prout H, Hood K, Kinnersley P. Management of suspected acute viral upper respiratory tract infection in children with intranasal sodium cromoglicate: a randomised controlled trial. Lancet 2002;359:2153-8.

20. Collet JP, Vanasse , Marois P, Amar M, Goldberg J, Lambert J, et al. Hyperbaric oxygen for children with cerebral palsy: a randomised multicentre trial. Lancet 2001;357:582-6.

21. Azzopardi DV, Strohm B, Edwards AD, Dyet L, Halliday HL, Juszczak E, et al. Moderate hypothermia to treat perinatal asphyxial encephalopathy. N Engl J Med 2009;361:1349-58.

22. Schmidt B, Roberts RS, Davis P, Doyle LW, Barrington KJ, Ohlsson A, et al. Long term effects of Caffeine Therapy for Apnea of Prematurity. N Engl J Med 2009;357(19):1893-1902.

23. Geller DA, Wagner KD, Emslie G, Murphy T, Carpenter D, Wetherhold E, et al. Paroxetine treatment in children and adolescents with obsessive-compulsive disorder: a randomized, multicenter, double-blind, placebo-controlled trial. J Am Acad Child Adolesc Psychiatry 2004;43:1387-96.

24. Riddle M, Reeve E, Yaryura-Tobias J, Yang HM, Claghorn JL, Gaffney G, et al. Fluvoxamine for children and adolescents with obsessive-compulsive disorder: a randomized, controlled, multicenter trial. J Am Acad Child Adolesc Psychiatry 2001;40:222-9.

25. van Vlimmeren LA, van der Graaf Y, Boere-Boonekamp M, L'Hoir MP, Helders PJM, Engelbert RHH. Effect of pediatric physical therapy on deformational plagiocephaly in children with positional preference. Arch Pediatr Adolesc Med 2008;162:712-8.

26. Lobaugh NJ, Karaskov V, Rombough V, Rovet J, Bryson S, Greenbaum R, et al. Piracetam therapy does not enhance cognitive functioning in children with down syndrome. Arch Pediatr Adolesc Med 2001;155:442-8.

27. Luria JW, Smith GA, Chapman JI. An evaluation of a safety education program for kindergarten and elementary school children. Arch Pediatr Adolesc Med 2000;154:227-31.

28. Daley AJ, Copeland RJ, Wright NP, Roalfe A, Wales JKH. Exercise therapy as a treatment for psychopathologic conditions in obese and morbidly obese adolescents: a randomized, controlled trial. Pediatrics 2006;118:2126-34.

29. Vagnoli L, Caprilli S, Robiglio A, Messeri A. Clown doctors as a treatment for preoperative anxiety in children: a randomized, prospective study. Pediatrics 2005;116:e563-7.

30. Shea S, Turgay A, Carroll A, Schulz M, Orlik H, Smith I, et al. Risperidone in the treatment of disruptive behavioral symptoms in children with autistic and other pervasive developmental disorders. Pediatrics 2004;114:e634-41.

31. Johnston CC, Filion F, Snider L, Majnemer An, Limperopoulos C, Walker CD, et al. Routine sucrose analgesia during the first week of life in neonates younger than 31 weeks' postconceptional age. Pediatrics 2002;110:523-8.