pISSN 2349-3240 | eISSN 2349-3259

DOI: http://dx.doi.org/10.18203/2349-3259.ijct20173133

Review Article

The impact of rater training on clinical outcomes assessment data: a literature review

Michael E. Sadler*, Rinah T. Yamamoto, Laura Khurana, Susan M. Dallabrida

eResearch Technology, 500 Rutherford Ave., Boston, MA, United States

Received: 19 May 2017
Revised: 29 June 2017
Accepted: 01 July 2017

***Correspondence:**
Dr. Michael E. Sadler,
E-mail: michael.sadler@ert.com

Copyright: © the author(s), publisher and licensee Medip Academy. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT

Rater training is a well-recognized approach to minimizing inaccuracy and variability in clinical outcomes assessments common in clinical trials. However, there is a dearth of empirical research on the types of rater training and qualifications that contribute to improved accuracy, inter-rater reliability and intra-rater reliability. Herein, we discuss the need for rater training in clinical trials and review publications that report data on rater characteristics, training modalities and outcomes in terms of accuracy and reliability of clinical outcomes data.

Keywords: Rater training, Clinical trials, Reliability, Accuracy

INTRODUCTION

Pharmaceutical clinical trials have increasingly grown in number, length, complexity and cost.[1] Ten to 12 years are now required to bring a new drug to market, with estimated total lifecycle costs per new compound (including cost of trial failures) of approximately 2.6 billion dollars. Of 1,442 compounds first tested in humans between 1995 and 2007 only 7.1% were approved and 80.3% were discontinued in some phase of clinical development.[2]

There are many potential contributors to inconclusive clinical trial results (failure to achieve primary endpoints), such as poor trial design, choice of endpoints or deficiencies in earlier phase programs.[3] However, often unreported are the quality and integrity of clinical outcomes assessments (COA).[4-6] Unlike quantitative biomarkers, with well-defined and precisely measureable values such as quantity or frequency, these outcome measures are subjective to varying degrees, and can be influenced by raters' judgments, training history and motivations.[5,7] The major rater-related challenges include consistency (inter- and intra-rater reliability) and accuracy (concordance with an expert rating or gold standard). In patient- and caregiver-reported outcomes ratings, individual's misunderstanding of concepts, terminology, scale and/or the role of the instrument in the trial often result in missing data and excessive variability.[8,9] These problems are especially acute in CNS areas, e.g., neurology and psychiatry where outcome measures such as semi-structured interviews rely heavily on clinical judgement.[10] For example, in depression, both the Montgomery-Åsberg Depression Rating Scale and Hamilton Rating Scale for Depression have been modified to improve inter- and intra-rater reliability.[11,12] However, COA in many other disciplines and indications are also vulnerable to human error, for example psoriasis, ophthalmology, Parkinson's disease, Alzheimer's disease, dermatology, rheumatology and multiple sclerosis.[13-23] In addition, patient reported outcomes (PROs) are commonly primary endpoints in pain, migraine, seizure, allergy, itch, and gastrointestinal diseases and as such, can be affected by the accuracy with which PRO data are captured.[21,24-26]

When the data from patient-reported, observer-reported and clinician-reported outcomes are primary outcome measures, they determine the success or failure of a trial.[11] If these assessments are inconsistent (unreliable) or of poor quality within and among raters, error variance will be high and the power to detect an effect low. If different raters use different criteria to complete the assessments (lack of standardization) or if raters' criteria change over the course of the trial (rater drift), the resulting data will be unreliable.[27-29] For example, trial failures are common in Alzheimer's drug development, many due to clinical rating "inaccuracies, imprecision, failures to follow or lack of operational protocols for applying methods, and bias."[10] Poor interview quality alone can introduce sufficient variability to yield inconclusive data.[6]

In addition to primary efficacy measures, supplementary assessments are used to measure other meaningful outcomes, such as quality of life or adverse events. For example, even in drug trials for skin disorders it has become critical to assess suicide ideation and behavior.[30] Thus, raters may be evaluating outcomes that are not within their areas of expertise and may be working with assessments for which they have no experience or training. This can degrade the reliability and validity of the outcome measurements if raters are not properly trained.

Globalization of trials has greatly increased the difficulty in monitoring and maintaining reliable data collection.[31,32] Clinical trials involving multiple sites, in multiple countries in many languages and cultures increase the need for well-trained and calibrated raters.[33,34] Rare diseases are also highly susceptible to problems of variability in outcomes measurements due to small subject numbers and diverse disease expression and patient experiences within the same condition.[35-37]

Rater training is a well-recognized approach to minimizing inaccuracies and variability in COA data, however there are currently no standards governing the selection of personnel for clinical raters who typically have widely varying types of training, levels of education and clinical experience.[27,38-40] Many COAs also rely on non-clinician observers (such as caregivers) and patients themselves, who may have difficulty understanding what is asked of them, as well as difficulties with compliance and reliability.[7] Regulatory and expert advice strongly endorse training for all raters of COAs including site raters, subjects, and caregivers.[9,41-43] Herein, we review the empirical evidence in support of rater training recommendations.

METHODS

Identification of eligible studies

We conducted a targeted search of the literature detailing the effects of rater training on clinical outcomes assessment with data on accuracy, inter-rater and/or intra-rater reliability. PubMed, ProQuest, Nursing & Allied Health, EBSCO, JSTOR and Web of Science databases were searched without restriction on publication dates. The first set of keywords used in a Boolean search of these databases was: rater training, inter-rater reliability, inter-rater agreement, rater education, investigator training, subject training, participant training, and caregiver training. A second set of keywords was: survey or questionnaire or instrument or patient reported outcome or clinical outcome assessment or scale. The two keyword searches were combined and filtered for the following terms in the title or abstract: humans, English and clinical trials (Table 1). Records were first screened by title and abstract prior to retrieving full-text articles for eligibility evaluation. The remaining articles were then hand-searched for additional citations (Table 1).

Publications were eligible if they described rater training on any clinical outcome measure in any therapeutic area and reported data on reliability and accuracy. We included prospective and retrospective studies. Non-English papers were excluded.

Table 1: Search strategy.

#	Search terms	PubMed	ProQuest Nursing and Allied Health	EBSCO	JSTOR	Web of science
1	Rater training or inter-rater reliability or inter-rater agreement or rater education or investigator training or subject training or participant training or caregiver training	85,307	208,755	16,995	114,183	2,617
2	Survey or questionnaire or instrument or patient reported outcome or clinical outcome assessment or scale	71,040	1,002,409	2,093,714	216,122	545,659
3	(#1 and #2) Humans, English and clinical Trials, in title/abstract	148	185	1396	4	158

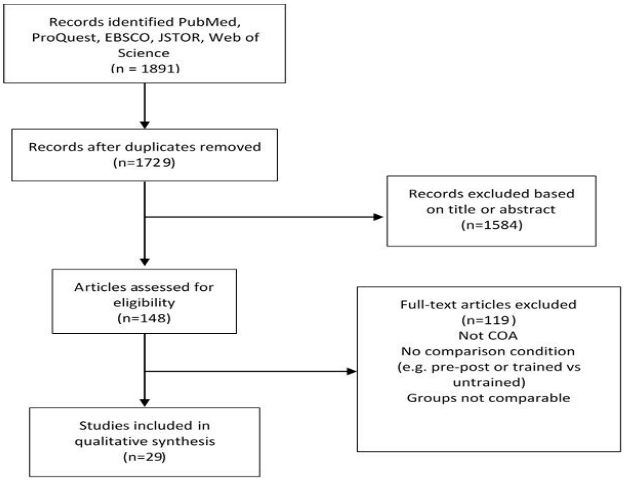
Records identified PubMed, ProQuest, EBSCO, JSTOR, Web of Science (n = 1891)
Records after duplicates removed (n=1729)
Records excluded based on title or abstract (n=1584)
Articles assessed for eligibility (n=148)
Full-text articles excluded (n=119) Not COA No comparison condition (e.g. pre-post or trained vs untrained) Groups not comparable
Studies included in qualitative synthesis (n=29)

Figure 1: Iterative reduction in literature process.

DISCUSSION

The initial searches yielded 427,857 articles for the first set of keywords and 3,928,944 for the second set. Combining the first two keyword searches with "and" and applying the filter terms, reduced the number of articles to 1891, some of which were research reports and others that were descriptive articles or summaries. All references were imported into EndNote software and duplicates removed for a total of 1725 articles (Table 1). The number of articles was further reduced through an iterative process (Figure 1). Study characteristics are summarized in Table 2.

Study characteristics

Disciplines, indications and instruments

Twenty-nine articles published between 1993 and 2016 met criteria for review. Psychiatry was the most common discipline with 14 studies across 5 indications and 6 assessment instruments. Depression and schizophrenia were the most common indications in psychiatry and the Hamilton Depression Scale (HAMD) and Positive and Negative Symptom Scales (PANSS) the most common instruments.[68,69] Five papers in neurology were identified, covering 5 indications and 5 instruments. Three articles concerned psoriasis and the Psoriasis Area Severity Index (PASI) Two studies concerned drug-induced movement disorders and the remaining 5 studies covered diverse medical indications and instruments (Table 2 and Figure 2).[70]

Training modalities

In this analysis, didactic refers to instructional material covering test administration and scoring, with or without discussion, delivered live or by video. Practical refers to scoring one or more interviews or other stimuli (video, audio, photographic or written) to a gold standard with feedback, with or without discussion. *Applied* training involves raters conducting and rating an actual interview, with live or remote observation and feedback on test administration and interviewing skills.

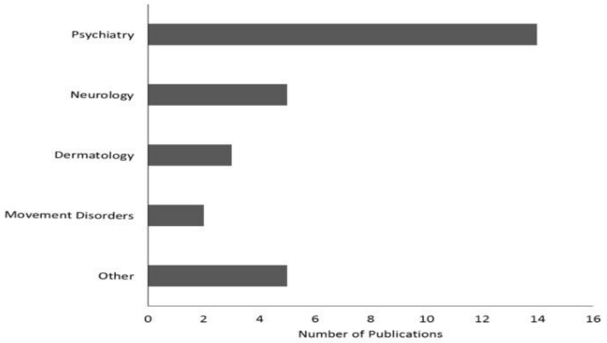

Figure 2: Distribution of publications by clinical research field.

The primary mode of training was didactic accompanied by some form of practice (n=18), applied training (n=4) or combined with a third mode (n=5) of instruction (Figure 3). Almost all studies employed some form of didactic, and most included a video demonstration of an administration of the instrument of interest, with actual patients or actors: "standardized patients." Methods for assessing the effects of training included comparison of pre- to post-training scores (n=14), training compared to no training (n=2) or sequential training (n=2). Many studies compared post-training rating scores to a "gold standard" or "expert consensus" score. Eleven studies did not include a comparison, reporting only reliability after training. Only 4 studies in psychiatry included applied assessment training. One study reported training of subjects, in addition to physicians, using the PASI for psoriasis.[20]

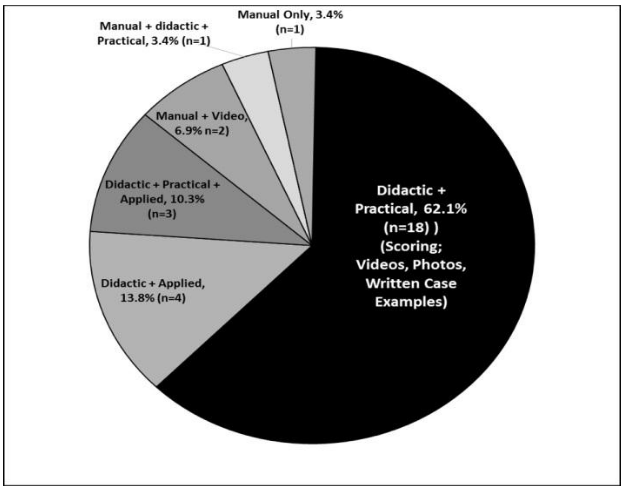

Figure 3: Percentage of studies using each training method.

Didactic= lecture (live or video) on rules for test administration, scoring, w/or without discussion; Practical = practice scoring interviews/stimuli (video, audio, photo or written) to a gold score w/feedback; Applied = conducting an interview and scoring, with live or remote observation and feedback on test administration and interviewing skills.

Table 2: Study characteristics.

Author	Date	Discipline	Indication	Instrument	n	Training manual	Didactic	Practical	Applied	Comparison (Pre/Post Training/No Training)	No difference	Significant improvement
Axelrod & Alphs[44]	1993	Psychiatry	Schizophrenia	NSA Negative	27		✓	✓				
Henrique-Araujo et al[45]	2014	Psychiatry	Depression	GRID HAMD	85	✓	✓	✓		✓	✓	
Jeglic et al[46]	2007	Psychiatry	Depression	HAMD HAMA	109		✓		✓	✓		✓
Kobak et al[47]	2003	Psychiatry	Depression	HAMD	9		✓	✓	✓	✓		✓
Kobak et al[48]	2007	Psychiatry	Schizophrenia	PANSS	12		✓	✓	✓	✓		✓
Kobak et al[38]	2005	Psychiatry	Depression	HAMD	46		✓		✓	✓		✓
Lundh et al[49]	2012	Psychiatry	Global impairment	CGAS	578		✓	✓		✓	✓	
Muller et al[50]	1998	Psychiatry	Schizophrenia	PANSS	23		✓	✓				
Muller & Wetzel[51]	1998	Psychiatry	Schizophrenia	PANSS	12		✓	✓				
Müller and Dragicevic[52]	2003	Psychiatry	Depression	HAMD	21		✓	✓				
Rosen et al[53]	2008	Psychiatry	Depression	GRID HAMD	13		✓	✓				
Tabuse et al[54]	2007	Psychiatry	Depression	GRID HAMD	70		✓	✓		✓	✓	
Targum[39]	2006	Psychiatry	Depression Anxiety Mania	HAMD HAMA YMRS	1241		✓		✓	✓		✓
Wagner et al[55]	2011	Psychiatry	Depression	HAMD IDS$_{C30}$	21		✓	✓				
Cusick et al[56]	2005	Neurology	Upper limb dysfunction	Melbourne Assessment	24	✓	✓	✓		✓	✓	
Kaufmann et al[57]	2007	Neurology	Amyotrophic Lateral Sclerosis	ALSFRS-R	76		✓		✓			
Russell et al. [58]	1994	Neurology	Cerebral palsy	GMFM	73		✓	✓		✓		✓
Schuld et al[59]	2013	Neurology	Spinal cord injury	ISNCSCI	106		✓	✓		✓		✓
Wilson et al[60]	2007	Neurology	Traumatic brain injury	Glasgow Outcome Scale	263		✓	✓		✓		✓
Armstrong et al[20]	2003	Dermatology	Psoriasis	PASI	56		✓	✓		✓		✓
Salvarani et al[19]	2016	Dermatology	Psoriasis	PASI	17		✓	✓		✓		✓
Youn et al[61]	2015	Dermatology	Psoriasis	PASI	21		✓	✓		✓		*

Inada et al[62]	1996	Movement disorders	Akathisia	Barnes Akathisia Scale	8		✓	✓		✓	✓
Loonen et al[63]	2001	Movement disorders	Drug-induced movement disorders	SADIMoD	6	✓					
Hansen et al.[64]	2015	Occupational therapy	Dysphagia	MISA	81		✓	✓	✓	✓	✓
Macnab et al[65]	1994	ICU	Sedative recovery	VSRS	16		✓	✓			
Prasad et al[16]	2015	Ophthalmology	Trachoma	WHO simplified trachoma grading system	8		✓	✓		✓	✓
Schaeffer[66]	2013	Speech language pathology	Dysphonia	DSP	5		✓				
Teal et al[67]	2012	Behavioral medicine	Diabetes	GET-D	7		✓	✓			

*One study saw mixed results; some improvement due to training, but not for every component of the assessment.

Assessment Abbreviations: NSA=Negative Symptom Assessment Scale; GRID-HAMD=GRID Hamilton Rating Scale for Depression; HAMD=Hamilton Rating Scale for Depression; HAMA-Hamilton Rating Scale for Anxiety; PANSS=Positive and Negative Symptom Scale; YMRS=Young Mania Rating Scale; IDS_{C30}=Inventory of Depressive Symptoms; ALSFRS-R=Amyotrophic Lateral Sclerosis Functional Rating Scale; GMFM=Gross Motor Function Measure; ISNCSCI=International Standards for Neurological Classification of Spinal Cord Injury; PASI=Psoriasis Area and Severity Index; SADIMoD=Schedule for the Assessment of Drug-Induced Movement Disorders; MISA=McGill Ingestive Skills Assessment; VSRS=Vancouver Sedative Recovery Scale; DSP=Dysphonic Severity Percentage Scale; GET-D=Goal-Setting Evaluation Tool for Diabetes

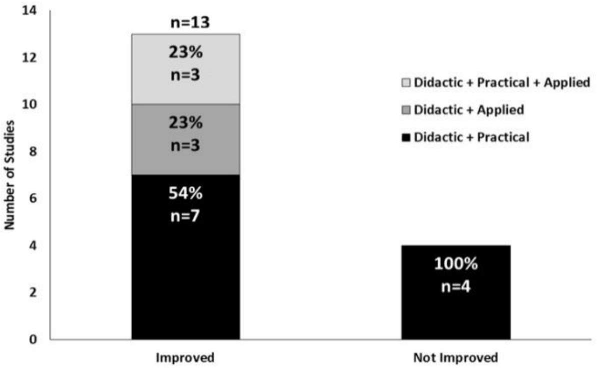

Figure 4: Effects of training methods by percentage of studies showing Improvement or No Difference (only studies in which a comparison was made either pre/post training or training/no training).

Effects of training

Studies that included a comparator were separated into those demonstrating statistically significant improvement versus no differences between conditions. Of the 13 studies that demonstrated improvement 7 used didactic instruction with practice, 3 employed didactic instruction with applied assessment training and 3 studies, using the HAM-D or PANSS included all 3 modalities. Four studies did not show improvements due to training; all 4 included didactic instruction with practice (Figure 4). In the 11 studies that did not include a comparison method, it was not possible to determine whether training improved rater skills. One study using didactic instruction with practice, demonstrated mixed results with some aspects improving while others did not.[61]

CONCLUSION

There is significant improvement in the accuracy and reliability of COAs across diverse indications when training meets certain standards. The following conclusions are supported:

- Without training, even experienced clinicians disagree on scoring and errors are common.[60,71] In some studies, training improved reliability more for raters with less experience.[19,45] Overall, however, rater training was effective regardless of discipline, education level, credentials or clinical experience.[38,39,53,54,59,72]
- COA didactic instruction should provide clear anchor points and objectively rated criteria for each item on an instrument, with coverage across all possible scores on each item.[44,53,58]
- Our findings are consonant with the Clinical Neuroscience Society (CNS) 2015 summit workshop recommendations for standards of rater training and demonstration of competence.[73] The proposed guidelines cover training and documentation for naïve and experienced raters, minimum standards for training and demonstration of competence, retraining over time, as well as multinational considerations for language and culture. CNS recommends didactic training that covers the purpose of the outcome assessment, standardization of administration, interview technique and scoring, as well as assessment of the raters' skills through practical training.
- Training programs that improved data quality were more comprehensive than those that were less effective or not effective. Furthermore, these programs were more intensive and of longer duration than what is typically conducted at investigator meetings.[38]
- In the training programs that did not demonstrate significant improvements in inter-rater reliability, 3 out of the 4 studies either had high ICCs before training or high ICCs in all groups. Some of these studies suffered from methodological issues. For example, in one study trainees watched 4 videotaped interviews in which the interviewers used a different version of the assessment being trained on. Another study used subjects who were all professionals with clinical experience and an assessment that is not considered to be difficult.
- There is clearly a place for clinical application as a component of rater training, particularly for interview skills.[6] However, it is difficult to draw conclusions from the limited number of studies in this review.

Reliability and accuracy of outcome measures in clinical research are essential for determining treatment efficacy in clinical trials. Considering the significant financial and medical stakes involved in clinical trial outcomes, it is critical and cost-effective to ensure raters are adequately trained.[74] Historically it has been well-recognized that clinical trials in the fields of psychiatry and neurology are especially vulnerable to rater error and bias due to the subjective nature of the COA.[29] It has been shown that poor interview quality on depression rating scales is directly related to trial failures.[5,6,29] For example, using data from drug trial of Paroxetine for depression, Kobak et al,[6] separated raters whose interviews were scored "good" or "excellent" using the RAPS for interview quality on the Hamilton Depression Rating Scale (HDRS).[75] When all the interviews were included regardless of interview quality, the results were not statistically significant, i.e., active treatment failed to separate from placebo. However, when only those interviews rated as "good" or "excellent" were analyzed the mean difference between the drug and placebo groups increased from 0.5 points on the HDRS to 6.83 points, resulting in a statistically significant effect in the drug group, with an effect size of 1.33. In this case, interview quality made the difference between a negative drug trial and a positive one.

Failure to separate active treatment from placebo is also common in schizophrenia trials. Khan, et al provided evidence of the relationship between low ratings reliability and failure in a schizophrenia trial with the clinician-rated PANSS as primary endpoint.[76] By partitioning the error variance into rater, subject, and time-points (number of visits), the authors showed that the source of unreliability was primarily found with raters for the placebo responders group. Placebo response is known to be particularly high in psychiatric trials and has been attributed to unreliable assessments by site raters.[77] However, our findings show that assessments in diverse disciplines benefit from rater training, as so many primary outcome measures rely on the interview skills and proficiency of site personnel.

There is a clear statistical relationship between the inter-rater reliability of outcome measures and the optimal number of subjects required to determine treatment efficacy.[78] As inter-rater reliability decreases, variability is introduced into the outcome being measured, resulting in greater difficulty separating the true measurement signal from error variance.[6] Therefore, significantly more subjects are required to determine a true effect. For example, if ICC decreases from 0.90 to 0.70, the power to detect an effect decreases from 0.72 to 0.50 and the number of subjects required to compensate increases by 22%.[74] Effective rater training can significantly reduce the number of subjects needed to power clinical trials, significantly reducing the cost to bring a drug to market.[79]

Limitations

Critical evaluation of these studies suggests that, aside from well-designed studies, many failed to use rigorous research methods to assess efficacy of training. We could not perform quantitative analyses because the information needed to assess effect sizes was not reported. A number of studies reported IRR as percent agreement, which is a poor determinant of reliability because it does not account for chance variability, while other studies reported a measure of IRR without reporting the specific statistic used.[80] Several studies suffered from small sample sizes and many studies in this review failed to provide comparative evidence for improvement from pre-training to post-training or active training compared to no training. Additionally, for a few studies where comparisons were made but results were not significant, it was difficult to discern whether the lack of results were due to inadequate training procedures or in some cases, ceiling effects. Another consideration that was difficult to address based on the literature was the appropriateness of training procedures for any given outcome measure, due to the limited details provided. Of the training methods described, didactic instruction along with practical and/or applied skills training appears to be the most common and most effective.

Recommendations

The findings of this review support recommendations for rater training across diverse indications to reduce variability in administration and scoring and mitigate failures to detect separation of active treatment from placebo. Where relevant, training should go beyond passive didactic instruction to include training and verification of raters' clinical research interviewing skills using interactive methodologies.

Further research should be designed to assess clearly detailed rater training, by comparing different methods of didactic instruction alone and in combination with practical and applied skills training appropriate to various measures. Studies should use naïve raters and include a comparison group of raters who receive no training. Raters should be assessed before training to determine baseline performance and after training to assess changes due to training. We suggest that future studies focus solely on quantitative assessment of rater training with sufficient sample sizes to detect changes in variability and accuracy for a moderate effect size with power set to 80%. Different clinical trial fields and assessment types may benefit from different training styles or components. Thus, it is necessary to test rater training effects in many disciplines including various types of instruments and interviews.

Funding: No funding sources
Conflict of interest: None declared
Ethical approval: Not required

REFERENCES

1. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nature Biotechnol. 2014;32(1):40-51.
2. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. In: Cost of developing a new drug. Boston: Tufts Center for the Study of Drug Development, Tufts University School of Medicine; 2014: 30.
3. Sertkaya A, Wong HH, Jessup A, Beleche T. Key cost drivers of pharmaceutical clinical trials in the United States. Clin Trials. 2016;13(2):117-26.
4. Mulsant BH. Interrater reliability in clinical trials of depressive disorders. Am J Psychiatry. 2002;159(9):1598-600.
5. Kobak KA, Kane JM, Thase ME, Nierenberg AA. Why do clinical trials fail? The problem of measurement error in clinical trials: Time to test new paradigms? J Clin Psychopharmacol. 2007;27(1):1-5.
6. Kobak KA, Feiger A, Lipsitz JD. Interview quality and signal detection in clinical trials. Am J Psychiatry. 2005;162(3):628.

7. Walton MK, Powers JH 3rd, Hobart J, Patrick D, Marquis P, Vamvakas S, et al. Clinical outcome assessments: Conceptual foundation-report of the ispor clinical outcomes assessment - emerging good practices for outcomes research task force. Value Health. 2015,18(6):741-52.
8. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. Health Technol Assess. 1998;2(14):i-iv,1-74.
9. FDA. Guidance for industry patient-reported outcome measures: Use in medical product development to support labeling claims MD: FDA. Available at: https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf Accessed on 3 March 2017.
10. Becker RE, Greig NH, Giacobini E. Why do so many drugs for alzheimer's disease fail in development? Time for new methods and new practices? J Alzheimers Dis. 2008;15(2):303-25.
11. Williams JB, Kobak KA. Development and reliability of a structured interview guide for the montgomery-asberg depression rating scale. Br J Psychiatry. 2008;192:52-8.
12. Williams JB. A structured interview guide for the Hamilton Depression Rating Scale. Arch Gen Psychiatry. 1988;45(8):742-7.
13. Berth-Jones J, Grotzinger K, Rainville C, Pham B, Huang J, Daly S, et al. A study examining inter- and intrarater reliability of three scales for measuring severity of psoriasis: Psoriasis Area and Severity Index, physician's global assessment and lattice system physician's global assessment. Br J Dermatol. 2006;155(4):707-13.
14. Puzenat E, Bronsard V, Prey S, Gourraud PA, Aractingi S, Bagot M et al. What are the best outcome measures for assessing plaque psoriasis severity? A systematic review of the literature. J Eur Acad Dermatol Venereol. 2010;24 Suppl 2:10-6.
15. Spuls PI, Lecluse LL, Poulsen ML, Bos JD, Stern RS, Nijsten T. How good are clinical severity and outcome measures for psoriasis? Quantitative evaluation in a systematic review. J Invest Dermatol 2010;130(4):933-43.
16. 16. Prasad BP, Bhatta RC, Chaudhary J, Sharma S, Mishra S, Cuddapah PA, et al. Agreement between novice and experienced trachoma graders improves after a single day of didactic training. Br J Ophthalmol. 2015;100(6):762-765.
17. Ramaker C, Marinus J, Stiggelbout AM, Van Hilten BJ. Systematic evaluation of rating scales for impairment and disability in parkinson's disease. Mov Disord. 2002;17(5):867-76.
18. Colell MG-V, March J, Sedway J. Rater qualifications in early alzheimer’s disease clinical trials. In: Alzheimer's & Dementia. vol. 10: Elsevier; 2014:4-178.
19. Salvarani C, Girolomoni G, Di Lernia V, Gisondi P, Tripepi G, Egan CG, et al. Impact of training on concordance among rheumatologists and dermatologists in the assessment of patients with psoriasis and psoriatic arthritis. Semin Arthritis Rheum. 2016;46(3):305-11.
20. Armstrong AW, Parsi K, Schupp CW, Mease PJ, Duffin KC. Standardizing training for psoriasis measures: Effectiveness of an online training video on psoriasis area and severity index assessment by physician and patient raters. JAMA Dermatol. 2013;149(5):577-82.
21. Charman C, Chambers C, Williams H. Measuring atopic dermatitis severity in randomized controlled clinical trials: What exactly are we measuring? J Invest Dermatol. 2003;120(6):932-41.
22. Pincus T. Limitations of a quantitative swollen and tender joint count to assess and monitor patients with rheumatoid arthritis. Bull NYU Hosp Jt Dis 2008;66(3):216-23.
23. Rudick RA, Larocca N, Hudson LD, Msoac. Multiple sclerosis outcome assessments consortium: Genesis and initial project plan. Mult Scler. 2014;20(1):12-7.
24. Dias N, Durand E, Gary S, Tuller J, Dallabrida S. Patients with gastrointestinal disorders prefer electronic and interactive training when participating in a clinical trial. In: DIA. Chicago, IL; 2017.
25. Dias N, Zhao L, Durand E, Gary S, Tuller J, Dallabrida S. Errors in patient reported outcomes (pros): Patients' understanding of how to record a headache day. In: ISPOR 22nd Annual International Meeting. Boston, MA; 2017.
26. Yamamoto R, Durand E, Gary S, Tuller J, Dallabrida S. Patient reported outcomes (pros) are subject to interpretation errors: Patients' understanding of how to report pain severity over a period of time. In: ISPOR 22nd Annual International Meeting. Boston, MA; 2017.
27. Kobak KA, Lipsitz JD, Williams JB, Engelhardt N, Jeglic E, Bellew KM. Are the effects of rater training sustainable? Results from a multicenter clinical trial. J Clin Psychopharmacol. 2007;27(5):534-6.
28. Rothman B, Yavorsky C, De Fries A, Gordon J, Opler M. P02-88 - quantifying rater drift on the ham-d in a sample of standardized rater training events: Implications for reliability and sample size calculations. European Psychiatry 2011;26:683.
29. Engelhardt N, Feiger AD, Cogger KO, Sikich D, DeBrota DJ, Lipsitz JD, et al. Rating the raters: Assessing the quality of hamilton rating scale for depression clinical interviews in two industry-sponsored clinical drug trials. J Clin Psychopharmacol. 2006;26(1):71-4.
30. Markus KA. FDA briefing document dermatologic and ophthalmic drugs advisory committee meeting. Silver Spring, MD: U.S. Food and Drug Administration, 2016.
31. English R, Lebovitz Y, Griffin R. Transforming clinical research in the United States: Challenges

and opportunities: Workshop summary. In: Forum on Drug Discovery, Development, and Translation; Institute of Medicine. Washington DC: National Academies Press (US); 2010.

32. Keefe RSE, Harvey PD. Implementation considerations for multisite clinical trials with cognitive neuroscience tasks. Schizophrenia Bulletin 2008;34(4):656-63.
33. Miller J. Complex clinical trials are posing new challenges across the clinical supply chain. BioPharm International 2010;23(4).
34. Small GW, Schneider LS, Hamilton SH, Bystritsky A, Meyers BS, Nemeroff C. Site variability in a multisite geriatric clinical trial. Int J Geriatric Psychiatry. 1996;11:1089-95.
35. Delaney KA. Tools to standardize assessments across multi-site trials: Methods to improve standardization of neuropsychological assessment in clinical trials. Maryland: U.S. Food and Drug Administration, 2015.
36. Pariser A. Rare disease and clinical trials. In. Edited by Administration USFaD. Maryland; 2014: 30.
37. FDA. Advancing the development of pediatric therapeutics workshop. Silver Springs, Maryland, 2015.
38. Kobak KA, Lipsitz JD, Williams JB, Engelhardt N, Bellew KM. A new approach to rater training and certification in a multicenter clinical trial. J Clin Psychopharmacol 2005;25(5):407-12.
39. Targum SD. Evaluating rater competency for cns clinical trials. J Clin Psychopharmacol. 2006;26(3):308-10.
40. Daniel D, Opler MGA, Wise-Rankovic A, Kalali A. Consensus recommendations on rater training and certification. In.: CNS Summit: Rater Training and Certification Workgroup; 2013:9.
41. EMA. Reflection paper on risk based quality management in clinical trials. UK; 2013. Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2013/11/WC500155491.pdf Accessed on 4 April 2017.
42. Kobak KA, Engelhardt N, Williams JB, Lipsitz JD. Rater training in multicenter clinical trials: Issues and recommendations. J Clin Psychopharmacol. 2004;24(2):113-7.
43. West MD, Daniel DG, Opler M, Wise-Rankovic A, Kalali A. Consensus recommendations on rater training and certification. Innov Clin Neurosci. 2014;11(11-12):10-3.
44. Axelrod BN, Alphs LD. Training novice raters on the negative symptom assessment scale. Schizophr Res. 1993;9(1):25-8.
45. Henrique-Araujo R, Osorio FL, Goncalves Ribeiro M, Soares Monteiro I, Williams JB, Kalali A, et al. Transcultural adaptation of grid hamilton rating scale for depression (grid-hamd) to brazilian portuguese and evaluation of the impact of training upon inter-rater reliability. Innov Clin Neurosci. 2014;11(7-8):10-8.
46. Jeglic E, Kobak KA, Engelhardt N, Williams JB, Lipsitz JD, Salvucci D, et al. A novel approach to rater training and certification in multinational trials. Int Clin Psychopharmacol. 2007;22(4):187-91.
47. Kobak KA, Lipsitz JD, Feiger A. Development of a standardized training program for the hamilton depression scale using internet-based technologies: Results from a pilot study. J Psychiatr Res. 2003;37(6):509.
48. Kobak KA, Opler MGA, Engelhardt N. Panss rater training using internet and videoconference: Results from a pilot study. Schizophr Res. 2007;92(1-3):63-7.
49. Lundh A, Kowalski J, Sundberg CJ, Landen M. A comparison of seminar and computer based training on the accuracy and reliability of raters using the children's global assessment scale (CGAS). Adm Policy Ment Health. 2012;39(6):458-65.
50. Müller MJ, Rossbach W, Dannigkeit P, Muller-Siecheneder F, Szegedi A, Wetzel H. Evaluation of standardized rater training for the positive and negative syndrome scale (PANSS). Schizophr Res. 1998;32(3):151-60.
51. Müller MJ, Wetzel H. Improvement of inter-rater reliability of PANSS items and subscales by a standardized rater training. Acta Psychiatr Scand 1998;98(2):135-9.
52. Müller MJ, Dragicevic A. Standardized rater training for the hamilton depression rating scale (HAMD-17) in psychiatric novices. J Affect Disord. 2003;77(1):65.
53. Rosen J, Mulsant BH, Marino P, Groening C, Young RC, Fox D. Web-based training and interrater reliability testing for scoring the hamilton depression rating scale. Psychiatry Res. 2008;161(1):126-30.
54. Tabuse H, Kalali A, Azuma H, Ozaki N, Iwata N, Naitoh H, et al. The new grid hamilton rating scale for depression demonstrates excellent inter-rater reliability for inexperienced and experienced raters before and after training. Psychiatry Res. 2007;153(1):61-7.
55. Wagner S, Helmreich I, Lieb K, Tadic A. Standardized rater training for the hamilton depression rating scale (HAMD(17)) and the inventory of depressive symptoms (IDSC30). Psychopathology. 2011;44(1):68-70.
56. Cusick A, Vasquez M, Knowles L, Wallen M. Effect of rater training on reliability of melbourne assessment of unilateral upper limb function scores. Development Med Child Neurol. 2005;47(1):39-45.
57. Kaufmann P, Levy G, Montes J, Buchsbaum R, Barsdorf AI, Battista V, et al. Excellent inter-rater, intra-rater, and telephone-administered reliability of the alsfrs-r in a multicenter clinical trial. Amyotroph Lateral Scler. 2007;8(1):42-6.
58. Russell DJ, Rosenbaum PL, Lane M, Gowland C, Goldsmith CH, Boyce WF, et al. Training users in the gross motor function measure: Methodological

and practical issues. Physical Therapy. 1994;74(7):630-6.

59. Schuld C, Wiese J, Franz S, Putz C, Stierle I, Smoor I, et al. Effect of formal training in scaling, scoring and classification of the international standards for neurological classification of spinal cord injury. Spinal Cord. 2013;51(4):282-8.
60. Wilson JT, Slieker FJ, Legrand V, Murray G, Stocchetti N, Maas AI. Observer variation in the assessment of outcome in traumatic brain injury: Experience from a multicenter, international randomized clinical trial. Neurosurgery. 2007;61(1):123-8.
61. Youn SW, Choi CW, Kim BR, Chae JB. Reduction of inter-rater and intra-rater variability in psoriasis area and severity index assessment by photographic training. Ann Dermatol. 2015;27(5):557-62.
62. Inada T, Matsuda G, Kitao Y, Nakamura A, Miyata R, Inagaki A, et al. Barnes Akathisia Scale: Usefulness of standardized videotape method in evaluation of the reliability and in training raters. Int J Methods Psychiatr Res. 1996;6(1):49-52.
63. Loonen AJ, Doorschot CH, van Hemert DA, Oostelbos MC, Sijben AE. The schedule for the assessment of drug-induced movement disorders (sadimod): Inter-rater reliability and construct validity. Int J Neuropsychopharmacol. 2001;4(4):347-60.
64. Hansen T, Elholm Madsen E, Sørensen A. The effect of rater training on scoring performance and scale-specific expertise amongst occupational therapists participating in a multicentre study: A single-group pre-post-test study. Disabil Rehabil 2015;38(12):1216-26.
65. Macnab AJ, Levine M, Glick N, Phillips N, Susak L, Elliott M. The Vancouver Sedative Recovery Scale for Children: Validation and reliability of scoring based on videotaped instruction. Can J Anaesth. 1994;41(10):913-8.
66. Schaeffer N. Student training to perceptually assess severity of dysphonia using the dysphonic severity percentage scale. J Voice. 2013;27(5):611-6.
67. Teal CR, Haidet P, Balasubramanyam AS, Rodriguez E, Naik AD. Measuring the quality of patients' goals and action plans: Development and validation of a novel tool. BMC Medical Info Decision Making. 2012;12(1):152-9.
68. Williams JB. A structured interview guide for the hamilton depression rating scale. Arch General Psychiatry. 1988;45(8):742-7.
69. 69. Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr Bull. 1987;13(2):261-76.
70. Fredriksson T, Pettersson U. Severe psoriasis--oral therapy with a new retinoid. Dermatologica. 1978;157(4):238-44.
71. Hilsabeck RC, Nations KR, Tanenbaum R, Grubb B, Choudhry A. Inter-rater reliability and error analysis of the scales for outcomes of parkinson’s disease: Cognition (scopa-cog) in moderato–a randomized double blind placebo controlled study to assess the effect of rasagiline on mild cognitive impairment in pd. In: Alzheimer's and Dementia. vol. 10; 2014: 854.
72. Busner J, Kott A, Sachs G. Increasing signal over noise in mdd clinical trials: Improvement after efficacy scale rater training among experienced mdd investigators. Eur Neuropsychopharmacology. 2013;23:348-8.
73. West MD, Daniel DG, Opler M, Wise-Rankovic A, Kalali A. Consensus recommendations on rater training and certification. Innov Clin Neurosci. 2015;11(11-12):10-3.
74. Perkins DO, Wyatt RJ, Bartko JJ. Penny-wise and pound-foolish: The impact of measurement error on sample size requirements in clinical trials. Biol Psychiatry. 2000;47(8):762-6.
75. Lipsitz J, Kobak K, Feiger A, Sikich D, Moroz G, Engelhard A. The rater applied performance scale: Development and reliability. Psychiatry Res. 2004;127(1-2):147-55.
76. Khan A, Yavorsky WC, Liechti S, DiClemente G, Rothman B, Opler M et al. Assessing the sources of unreliability (rater, subject, time-point) in a failed clinical trial using items of the positive and negative syndrome scale (PANSS). J Clin Psychopharmacol. 2013;33(1):109-17.
77. Kinon BJ, Potts AJ, Watson SB. Placebo response in clinical trials with schizophrenia patients. Curr Opin Psychiatry. 2011;24(2):107-13.
78. Gwet KL. Handbook of inter-rater reliability, 4th edition: The definitive guide to measuring the extent of agreement among raters Maryland: Advanced Analytics, LLC; 2014.
79. Leon AC. Implications of clinical trial design on sample size requirements. Schizophr Bull. 2008;34(4):664-9.
80. Hallgren KA. Computing inter-rater reliability for observational data: An overview and tutorial. Tutor Quant Methods Psychol. 2012;8(1):23-34.

Cite this article as: Sadler ME, Yamamoto RT, Khurana L, Dallabrida SM. The impact of rater training on clinical outcomes assessment data: a literature review. Int J Clin Trials 2017;4(3):101-10.