THE IMPACT OF RATER TRAINING ON CLINICAL OUTCOMES ASSESSMENT DATA: A CONCEPTUAL REVIEW

Sadler ME[1], Yamamoto RT[1], Khurana L[1], Dallabrida SM[1]

[1]eResearch Technology

INTRODUCTION

In the current context of escalating drug development costs, clinical trial length and complexity and frequent trial failure, data integrity is critical[1]. Much of the primary and secondary endpoint data collected across indications utilizes various clinical outcomes assessments (COAs). However, the methods for ensuring the accuracy and reliability of COAs in clinical trials are often unreported. Training of those individuals (raters) who complete COAs, including site personnel, patients, and caregivers is recommended by experts in the field, and standards have been proposed for training, evaluation and monitoring of rater performance. Regulatory authorities increasingly seek evidence of training in the evaluation of labelling claims based on COA data.

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, misunderstanding of concepts, terminology, scale and/or the role of the instrument in the trial often result in missing data and excessive variability[2,3]. In CNS areas, e.g., neurology and psychiatry, outcome measures such as semi-structured interviews rely heavily on clinical judgement[4] and outcomes assessments in many other therapeutic areas are also vulnerable to human error, for example psoriasis[5-7], ophthalmology[8], Parkinson's disease[9], Alzheimer's disease[10], dermatology[11,12] and multiple sclerosis[13].

Globalization of trials has greatly increased the difficulty in monitoring and maintaining reliable data collection[14,15]. Clinical trials involving multiple sites, in multiple countries in many languages and cultures[16] increase the need for well-trained and calibrated raters[17]. This is even more critical in clinical trials for rare diseases. Rare diseases, in addition to the global challenges, are highly susceptible to problems of variability in outcomes measurements due to small subject numbers, diverse disease expression and patient experiences within the same condition[18-20].

This review aimed to evaluate the empirical evidence for the impact of rater training on the quality of clinical outcomes assessments (COAs).

METHODS

A targeted review of 1,891 articles published 1993–2016 identified 29 eligible papers containing data on rater training and the reliability and accuracy of COA data applicable to clinical trials in any therapeutic area. Non-English papers were excluded (Figure 1). Included studies reported training of site staff/clinicians; only 1 study reported effects of patient training.

RESULTS

We classified rater training elements as didactic (video or live lecture on administration and scoring with or without discussion), practical (practice scoring of interviews or other stimuli to a "gold standard" with feedback, with or without discussion), and clinical (conducting and scoring an interview with live or remote observation and feedback on test administration and clinical interviewing skills, with remediation if needed).

Most studies were in psychiatry (n=14) for depression using the Hamilton Depression Scale and schizophrenia using the Positive and Negative Symptom Scales, followed by neurology (n=5), psoriasis (n=3), drug-induced movement disorders (n=2), and miscellaneous (n=5) (Figure 2). All studies reported some form of didactic instruction, most accompanied by practical training (n=18) and/or clinical training (n=4). The study characteristics are summarized in Table 1.

Of the studies that included a comparison group 13 showed improvement while 4 found no differences with training. In the 11 studies lacking a comparison group it was not possible to determine whether training improved rater skills or data quality due to a lack of relevant data (Table 1 and Figure 3).

Table 1: Study characteristics

Author	Date	Discipline	Indication	Instrument	n	Training manual	Didactic	Practical	Clinical	Comparison (Pre/Post Training/No	No difference	Significant Improvement
Axelrod & Alphs	(1993)	Psychiatry	Schizophrenia	NSA	27		■	■				
Henrique-Araujo et al.	(2014)	Psychiatry	Depression	GRID HAMD	85	■	■	■		■	■	
Jeglic et al.	(2007)	Psychiatry	Depression	HAMD HAMA	109		■		■	■		■
Kobak et al.	(2003)	Psychiatry	Depression	HAMD	9		■	■	■	■		■
Kobak et al.	(2007c)	Psychiatry	Schizophrenia	PANSS	12		■	■	■	■		■
Kobak et al.	(2005b)	Psychiatry	Depression	HAMD	46		■		■	■		■
Lundh et al.	(2012)	Psychiatry	Global impairment	CGAS	578		■	■		■	■	
Muller et al.	(1998)	Psychiatry	Schizophrenia	PANSS	23		■	■				
Muller & Wetzel	(1998)	Psychiatry	Schizophrenia	PANSS	12		■	■				
Müller and Dragicevic	(2003)	Psychiatry	Depression	HAMD	21		■	■				
Rosen et al.	(2008)	Psychiatry	Depression	GRID HAMD	13		■	■				
Tabuse et al.	(2007)	Psychiatry	Depression	GRID HAMD	70		■	■		■	■	
Targum	(2006)	Psychiatry	Depression Anxiety Mania	HAMD HAMA YMRS			■		■	■		■
Wagner et al.	(2011)	Psychiatry	Depression	HAMD IDS_{C30}	21		■	■				
Cusick et al.	(2005)	Neurology	Upper limb dysfunction	Melbourne Assessment	24	■	■	■		■	■	
Kaufmann et al.	(2007)	Neurology	Amyotrophic Lateral Sclerosis	ALSFRS-R	76		■		■			
Russell et al.	(1994)	Neurology	Cerebral palsy	GMFM	73		■	■		■		■
Schuld et al.	(2013)	Neurology	Spinal cord injury	ISNCSCI	106		■	■		■		■
Wilson et al.	(2007)	Neurology	Traumatic brain injury	Glasgow Outcome Scale	263		■	■		■		■
Armstrong et al.	(2013)	Dermatology	Psoriasis	PASI	56		■	■		■		■
Salvarani et al.	2(2016)	Dermatology	Psoriasis	PASI	17		■	■		■		■
Youn et al.	(2015)	Dermatology	Psoriasis	PASI	21		■	■		■		'
Inada et al.	(1996)	Movement disorders	Akathisia	Barnes Akathisia Scale	8		■	■		■		■
Loonen et al.	(2001)	Movement disorders	Drug-induced movement disorders	SADIMoD	6	■						
Hansen et al.	(2015)	Occupational therapy	Dysphagia	MISA	81		■	■	■	■		■
Macnab et al.	(1994)	ICU	Sedative recovery	VSRS	16		■	■				
Prasad et al.	(2015)	Ophthalmology	Trachoma	WHO simplified trachoma grading system	8		■	■		■		■
Schaeffer	(2013)	Speech-language pathology	Dysphonia	DSP	5		■					
Teal et al.	(2012)	Behavioral medicine	Diabetes	GET-D	7		■	■				

Figure 1: Flow diagram of iterative reduction in literature process

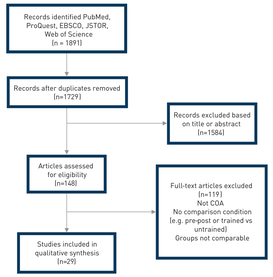


Figure 2. Distribution of articles by clinical research field.

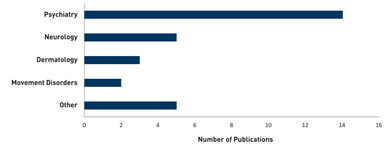
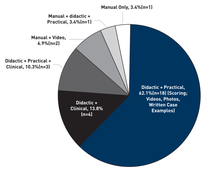
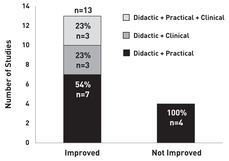


Figure 3. Percent of studies using each training method.



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
Clinical = conducting an interview and scoring, with live or remote observation and feedback on test administration and interviewing skills

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



CONCLUSION

The findings indicate that rater training is associated with significant improvement in the accuracy and reliability of COAs across diverse indications and instruments when training meets certain standards. The following conclusions are supported:

- Without training, even experienced clinicians disagree on scoring and errors are common[21,22]. In some studies, training improved reliability more for raters with less experience[11,23]. Overall, however, rater training was effective regardless of discipline, education level, credentials or clinical experience[24–29]
- Training on semi-structured interviews such as the Hamilton Depression Rating Scale and Positive and Negative Symptom Scales may benefit from a clinical component focusing on raters' ability to apply scoring criteria in an actual interview[30–32].
- Our findings are consonant with the Clinical Neuroscience Society (CNS) 2015 summit workshop recommendations for standards of rater training and demonstration of competence[33]. 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' interview skills.
- Training should provide clear anchor points and objective scoring criteria across the range of possible scores on each item of an instrument.
- 76% of studies that included a comparison group, demonstrated improved inter- and intra-rater reliability following any combination of training components (didactic, practical and clinical).
- 3 of the 4 studies that did not demonstrate significant improvements in inter-rater reliability had either high ICCs before training or high ICCs in all groups. Some of these studies suffered from methodological flaws. For example, in one study trainees watched 4 videotaped interviews in which the interviewers did not use the instrument being trained on. Another study used subjects who were all professionals with clinical experience and an assessment that is not considered to be difficult.

DISCLOSURES

The authors are employees of ERT.

REFERENCES

1. Hay, M., Thomas, D. W., Craighead, J. L. Economides, C., Rosenthal, J., Clinical development success rates for investigational drugs. Nature Biotechnology, 2014. 32(1).

2. Fitzpatrick, R., et al., Evaluating patient-based outcome measures for use in clinical trials. Health Technol Assess, 1998. 2(14): p. i-iv, 1-74.

3. FDA, Guidance for Industry Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims U.S.D.o.H.a.H. Services, Editor. 2009, FDA: MD.

4. Becker, R.E., N.H. Greig, and E. Giacobini, 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): p. 303-25.

5. Berth-Jones, J., 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): p. 707-13.

6. Puzenat, E., 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: p. 10-6.

7. Spuls, P.I., et al., How good are clinical severity and outcome measures for psoriasis?: quantitative evaluation in a systematic review. J Invest Dermatol, 2010. 130(4): p. 933-43.

8. Prasad, B.P., et al., Agreement between novice and experienced trachoma graders improves after a single day of didactic training. Br J Ophthalmol, 2015. 100(6): p. 762-5.

9. Ramaker, C., et al., Systematic evaluation of rating scales for impairment and disability in Parkinson's disease. Mov Disord, 2002. 17(5): p. 867-76.

10. Colell, M.G.-V., J. March, and J. Sedway, Rater qualifications in early Alzheimer's disease clinical trials, in Alzheimer's & Dementia. 2014, Elsevier: p. P4-178.

11. Salvarani, C., 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.

12. Armstrong, A.W., et al., 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): p. 577-82.

13. Rudick, R.A., et al., Multiple Sclerosis Outcome Assessments Consortium: Genesis and initial project plan. Mult Scler, 2014. 20(1): p. 12-7.

14. English, R., Y. Lebovitz, and R. Griffin, Transforming Clinical Research in the United States: Challenges and Opportunities: Workshop Summary, in Forum on Drug Discovery, Development, and Translation; Institute of Medicine. 2010, National Academies Press (US): Washington DC.

15. Keefe, R.S.E. and P.D. Harvey, Implementation Considerations for Multisite Clinical Trials with Cognitive Neuroscience Tasks. Schizophrenia Bulletin, 2008. 34(4): p. 656-663.

16. Miller, J., Complex Clinical Trials are Posing New Challenges Across the Clinical Supply Chain. BioPharm International, 2010. 23(4).

17. Small, G.W., Schneider, L.S, Hamilton, S H, Bystritsky, A, Meyers, B S, Nemeroff, C, Site variability in a multisite geriatric clinical trial. International Journal of Geriatric Psychiatry, 1996. 11: p. 1089-1095.

18. Delaney, K.A., Tools to Standardize Assessments Across Multi-Site Trials

Methods to Improve Standardization of Neuropsychological Assessment in Clinical Trials, in Assessing Neurocognitive Outcomes in Inborn Errors of Metabolism, U.S.F.a.D. Administration, Editor. 2015: Maryland. p. 42.

19. Pariser, A., Rare Disease and Clinical Trials, U.S.F.a.D. Administration, Editor. 2014: Maryland. p. 30.

20. FDA, Advancing the development of pediatric therapeutics workshop, U.S.F.a.D. Administration, Editor. 2015: Silver Springs, Maryland. p. 312.

21. Hilsabeck, R.C., et al., 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 & Dementia. 2014. p. P854.

22. Wilson, J.T., et al., Observer variation in the assessment of outcome in traumatic brain injury: experience from a multicenter, international randomized clinical trial. Neurosurgery, 2007. 61(1): p. 123-8; discussion 128-9.

23. Henrique-Araujo, R., 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): p. 10-8.

24. Kobak, K.A., et al., A new approach to rater training and certification in a multicenter clinical trial. J Clin Psychopharmacol, 2005. 25(5): p. 407-12.

25. Rosen, J., et al., Web-based training and interrater reliability testing for scoring the Hamilton Depression Rating Scale. Psychiatry Research, 2008. 161(1): p. 126-130.

26. Tabuse, 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 Research, 2007. 153(1): p. 61-67.

27. Targum, S.D., Evaluating rater competency for CNS clinical trials. J Clin Psychopharmacol, 2006. 26(3): p. 308-10.

28. Schuld, C., 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): p. 282-288.

29. Busner, J., A. Kott, and G. Sachs, Increasing signal over noise in MDD clinical trials: improvement after efficacy scale rater training among experienced MDD investigators. European Neuropsychopharmacology, 2013. 23: p. S348-S348.

30. Jeglic, E., et al., A novel approach to rater training and certification in multinational trials. Int Clin Psychopharmacol, 2007. 22(4): p. 187-91.

31. Kobak, K.A., et al., Are the effects of rater training sustainable? Results from a multicenter clinical trial. Journal of Clinical Psychopharmacology, 2007. 27(5): p. 534-536.

32. Kobak, K.A., A. Feiger, and J.D. Lipsitz, Interview quality and signal detection in clinical trials. American Journal of Psychiatry, 2005. 162(3): p. 628.

33. West, M.D., et al., Consensus recommendations on rater training and certification. Innov Clin Neurosci, 2015. 11(11-12): p. 10-3.

ERT®
Confidence at every turn™

Presented at the International Society for CNS Clinical Trials and Methodology, 21-23 February in Washington DC | For more information, contact: ert.com/contact-us

ert.com

This documentation and the data contained herein is the property of ERT, located at 1818 Market St., Suite 1000 Philadelphia, PA 19103, and its licensors. Any use, disclosure, reproduction, modification, display or transfer of the data and/or documentation is prohibited, except by the express written authorization of ERT. This document contains proprietary ERT information. Copyright © 1972-2016 by ERT. All rights reserved. ERT® is a trademark of ERT. All other trademarks are the property of their respective owners. THE INFORMATION IN THIS DOCUMENT IS SUBJECT TO CHANGE WITHOUT NOTICE. ERT MAKES NO WARRANTIES OF ANY KIND REGARDING THIS MATERIAL AND ASSUMES NO RESPONSIBILITY FOR ANY ERRORS THAT MAY APPEAR IN THIS DOCUMENT.