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

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. 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.

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. However, often unreported are the quality and integrity of clinical outcomes assessments (COA). Unlike quantitative biomarkers, with well-defined and precisely measurable 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. 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. 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. 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. However, COA in many other disciplines and indications are also vulnerable to human error, e.g. psoriasis, ophthalmology, Parkinson’s disease, Alzheimer’s disease, dermatology, rheumatology and multiple sclerosis. 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.

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.\textsuperscript{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.\textsuperscript{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.”\textsuperscript{10} Poor interview quality alone can introduce sufficient variability to yield inconclusive data.\textsuperscript{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.\textsuperscript{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.\textsuperscript{31,32} Clinical trials involving multiple sites, in multiple countries in many languages and cultures increase the need for well-trained and calibrated raters.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{7} Regulatory and expert advice strongly endorse training for all raters of COAs including site raters, subjects, and caregivers.\textsuperscript{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>
<thead>
<tr>
<th>#</th>
<th>Search terms</th>
<th>PubMed</th>
<th>ProQuest Nursing and Allied Health</th>
<th>EBSCO</th>
<th>JSTOR</th>
<th>Web of science</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>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</td>
<td>85,307</td>
<td>208,755</td>
<td>16,995</td>
<td>114,183</td>
<td>2,617</td>
</tr>
<tr>
<td>2</td>
<td>Survey or questionnaire or instrument or patient reported outcome or clinical outcome assessment or scale</td>
<td>71,040</td>
<td>1,002,409</td>
<td>2,093,714</td>
<td>216,122</td>
<td>545,659</td>
</tr>
<tr>
<td>3</td>
<td>(#1 and #2) Humans, English and clinical Trials, in title/abstract</td>
<td>148</td>
<td>185</td>
<td>1396</td>
<td>4</td>
<td>158</td>
</tr>
</tbody>
</table>

Table 1: Search strategy.
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. 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).

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.

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 1: Iterative reduction in literature process.

Figure 2: Distribution of publications by clinical research field.

Figure 3: Percentage of studies using each training method.

Didactic = lecture (live or video) on rules for test administration, scoring, with 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.

<table>
<thead>
<tr>
<th>Author</th>
<th>Date</th>
<th>Discipline</th>
<th>Indication</th>
<th>Instrument</th>
<th>n</th>
<th>Training manual</th>
<th>Didactic</th>
<th>Practical</th>
<th>Applied</th>
<th>Comparison (Pre/Post Training/No Training)</th>
<th>No difference</th>
<th>Significant Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Axelrod &amp; Alphs</td>
<td>1993</td>
<td>Psychiatry</td>
<td>Schizophrenia</td>
<td>NSA Negative</td>
<td>27</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Henrique-Araujo et al</td>
<td>2014</td>
<td>Psychiatry</td>
<td>Depression</td>
<td>GRID HAMD</td>
<td>85</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Jeglic et al</td>
<td>2007</td>
<td>Psychiatry</td>
<td>Depression</td>
<td>HAMD HAMA</td>
<td>109</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Kobak et al</td>
<td>2003</td>
<td>Psychiatry</td>
<td>Depression</td>
<td>HAMD</td>
<td>9</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Kobak et al</td>
<td>2007</td>
<td>Psychiatry</td>
<td>Schizophrenia</td>
<td>PANSS</td>
<td>12</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Kobak et al</td>
<td>2005</td>
<td>Psychiatry</td>
<td>Depression</td>
<td>HAMD</td>
<td>46</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Lundh et al</td>
<td>2012</td>
<td>Psychiatry</td>
<td>Global impairment</td>
<td>CGAS</td>
<td>578</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Muller et al</td>
<td>1998</td>
<td>Psychiatry</td>
<td>Schizophrenia</td>
<td>PANSS</td>
<td>23</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Muller &amp; Wetzel</td>
<td>1998</td>
<td>Psychiatry</td>
<td>Schizophrenia</td>
<td>PANSS</td>
<td>12</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Müller and Dragicevic</td>
<td>2003</td>
<td>Psychiatry</td>
<td>Depression</td>
<td>HAMD</td>
<td>21</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Rosen et al</td>
<td>2008</td>
<td>Psychiatry</td>
<td>Depression</td>
<td>GRID HAMD</td>
<td>13</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Tabuse et al</td>
<td>2007</td>
<td>Psychiatry</td>
<td>Depression</td>
<td>GRID HAMD</td>
<td>70</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Targum</td>
<td>2006</td>
<td>Psychiatry</td>
<td>Depression Anxiety Mania</td>
<td>HAMD HAMA YMRS</td>
<td>1241</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Wagner et al</td>
<td>2011</td>
<td>Psychiatry</td>
<td>Depression</td>
<td>HAMD IDS$_{30}$</td>
<td>21</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Cusick et al</td>
<td>2005</td>
<td>Neurology</td>
<td>Upper limb dysfunction</td>
<td>Melbourne Assessment</td>
<td>24</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Kaufmann et al</td>
<td>2007</td>
<td>Neurology</td>
<td>Amyotrophic Lateral Sclerosis</td>
<td>ALSFRS-R</td>
<td>76</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Russell et al</td>
<td>1994</td>
<td>Neurology</td>
<td>Cerebral palsy</td>
<td>GMFM</td>
<td>73</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Schuld et al</td>
<td>2013</td>
<td>Neurology</td>
<td>Spinal cord injury</td>
<td>ISNCS</td>
<td>106</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Wilson et al</td>
<td>2007</td>
<td>Neurology</td>
<td>Traumatic brain injury</td>
<td>Glasgow Outcome Scale</td>
<td>263</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Armstrong et al</td>
<td>2003</td>
<td>Dermatology</td>
<td>Psoriasis</td>
<td>PASI</td>
<td>56</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Salvarani et al</td>
<td>2016</td>
<td>Dermatology</td>
<td>Psoriasis</td>
<td>PASI</td>
<td>17</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Youn et al</td>
<td>2015</td>
<td>Dermatology</td>
<td>Psoriasis</td>
<td>PASI</td>
<td>21</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Authors</td>
<td>Year</td>
<td>Field</td>
<td>Condition/Procedure</td>
<td>Tool or Scale</td>
<td>Score</td>
<td>Components</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>------</td>
<td>-------------------------------</td>
<td>------------------------------------------</td>
<td>------------------------</td>
<td>-------</td>
<td>------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inada et al.</td>
<td>1996</td>
<td>Movement disorders</td>
<td>Akathisia</td>
<td>Barnes Akathisia Scale</td>
<td>8</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loonen et al.</td>
<td>2001</td>
<td>Movement disorders</td>
<td>Drug-induced movement disorders</td>
<td>SADIMoD</td>
<td>6</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hansen et al.</td>
<td>2015</td>
<td>Occupational therapy</td>
<td>Dysphagia</td>
<td>MISA</td>
<td>81</td>
<td>✓ ✓ ✓ ✓ ✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Macnab et al.</td>
<td>1994</td>
<td>ICU</td>
<td>Sedative recovery</td>
<td>VSRS</td>
<td>16</td>
<td>✓ ✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prasad et al.</td>
<td>2015</td>
<td>Ophthalmology</td>
<td>Trachoma</td>
<td>WHO simplified trachoma grading system</td>
<td>8</td>
<td>✓ ✓ ✓ ✓ ✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schaeffer</td>
<td>2013</td>
<td>Speech language pathology</td>
<td>Dysphonia</td>
<td>DSP</td>
<td>5</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teal et al.</td>
<td>2012</td>
<td>Behavioral medicine</td>
<td>Diabetes</td>
<td>GET-D</td>
<td>7</td>
<td>✓ ✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*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-HAM-A 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
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. 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. 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. 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. 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%. 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.

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. 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


37. FDA. Advancing the development of pediatric therapeutics workshop. Silver Springs, Maryland, 2015.


