INTRODUCTION
In the current context of escalating drug development costs, clinical trial length and complexity and frequent trial failure, data integrity is critical. 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, 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 labeling 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. In CNS areas, e.g., neurology and psychiatry, outcome measures such as semi-structured interviews rely heavily on clinical judgement and outcomes assessments in many other therapeutic areas are also vulnerable to human error, for example psoriasis, ophthalmology, Parkinson’s disease, Alzheimer’s disease, dermatology and multiple sclerosis.

Globalization of trials has greatly increased the difficulty in monitoring and maintaining reliable data collection. Clinical trials involving multiple sites, in multiple countries in many languages and cultures increase the need for well-trained and calibrated raters. 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.

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 or without discussion), practical (practical 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=46) for depression using the Hamilton Depression Scale and schizophrenia using the Positive and Negative Symptom Scales, followed by neurology (n=3), ophthalmology and urology (n=2), and miscellaneous (n=3) (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 18 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).

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 assessments, and this variability is common. In some studies, training improved reliability more for raters with less experience.

• Overall, however, rater training was effective regardless of discipline, education, level, credentials or clinical experience.

• 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.

• Our findings are consonant with the Clinical Neuroscience Society CNS 2015 summit workshop recommendations for standards of rater training and demonstration of competence. The proposed guidelines cover training and documentation for naive and experienced raters, minimum standards for training and demonstration of competence, retraining over time, as well as additional 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-rater 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 both 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 a 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