Original Research

The Journal of Prevention of Alzheimer's Disease - JPAD©
Volume 5, Number 3, 2018

Detection of Rater Errors on Cognitive Instruments in a Clinical Trial Setting

D.J. Connor[1], C.W. Jenkins[2], D. Carpenter[3], R. Crean[3], P. Perera[3]

1. Consultants in Cognitive and Clinical Trials, San Diego, CA, 92163, USA; 2. Alzheimer's Therapeutic Research Institute and Department of Neurology, Keck School of Medicine of University of Southern California, San Diego, CA, 92121 USA; 3. Dart Neuroscience, San Diego, CA, 92131, USA

Corresponding Author: Donald J. Connor PhD, PhD; Consultants in Cognitive and Clinical Trials; PO Box 33724; San Diego, CA; 92163, donconnorphd@yahoo.com, 623-332-5393

J Prev Alz Dis 2018;5(3):188-196
Published online May 22, 2018, http://dx.doi.org/10.14283/jpad.2018.20

Abstract

OBJECTIVES: This study examines errors committed by raters in a clinical trial of a memory enhancement compound.

BACKGROUND: Findings of clinical trials are directly dependent on the quality of the data obtained but there is little literature on rates or nature of rater errors on cognitive instruments in a multi-site setting.

DESIGN: Double-blind placebo-controlled study.

SETTING: 21 clinical sites in North America.

PARTICIPANTS: Two hundred seventy-five participants.

MEASUREMENTS: MMSE, WMS-R Logical Memory I & II, WMS-R Verbal Paired Associates I, WASi Vocabulary, WASi Matrix Reasoning, GDS and MAC-Q.

RESULTS: The WMS-R Logical Memory I & II and WASi Vocabulary tests were found to have the greatest number of scoring errors. Few substantive errors were detected on source document review of the MMSE, GDS, MAC-Q and WMS-R Verbal Paired Associates I. Some additional administration and scoring issues were identified during feedback sessions with the raters.

CONCLUSIONS: Cognitive measures used in clinical trials are prone to errors which can be detected with proper monitoring. Some instruments are particularly prone to inter-rater variably and should therefore be targets for focused training and ongoing monitoring. Areas in need of further investigation to help inform and optimize quality of clinical trial data are discussed.

Introduction

Conclusions drawn about treatment efficacy in clinical trials are directly dependent on the quality of assessment data. It is assumed that the cognitive instruments used in clinical trials are well validated in the population to be assessed, that they have good inter- and intra-rater reliability, and that raters administer and score them in accordance with a standard test protocol. In clinical trials multiple factors can and often do challenge these assumptions, potentially impacting validity and reliability of the data (Table 1). Instrumental factors may include psychometric characteristics of a selected instrument that are inappropriate to the study population, ambiguous administration or scoring instructions, failure to define appropriate and inappropriate prompts, cues, and corrections, or score sheets lacking key information to assist raters during test administration (e.g. verbatim instructions to subject, start/stop rules, proper delay intervals, etc.). Other factors that may affect quality of data are related to characteristics of the raters collecting it, including depth of their experience with the instruments and with the study population, knowledge of psychometric theory and factors influencing subject performance, proper trial-specific training and adequate qualification/certification procedures, rater willingness to seek help when needed, and availability of experts to provide such guidance. Inattention to these instrument- and rater-related factors during a clinical trial can contribute to rater errors in administration and scoring, leading ultimately to degradation in instruments' ability to detect meaningful change (1, 2).

Investigator training meetings must cover the overall study protocol, therefore discussion of specific administration and scoring guidelines for selected cognitive and clinical instruments can be quite limited in that venue. Methods of 'fine-tuning' technique, including illustration of common errors made by raters, common response deviations made by participants, and discussion of allowable vs inappropriate prompts, corrections and feedback are generally not part of the training agenda. Instrument administration manuals may be provided to guide additional independent learning by raters, but are of benefit only if sufficiently detailed, clear and referred to regularly by raters throughout the trial. Failure on any of these counts introduces risk of unwanted variability in collected data.

Another issue affecting rater reliability may be inconsistency of training on the same instruments being used in different clinical trials. Previous studies have highlighted this kind of variability in training of the same outcome measures across different clinical trials (3, 4). As study duration of clinical trials has increased in recent years, so has the likelihood of rater exposure to multiple trials using similar measures, and consequently the

Received December 12, 2017
Accepted for publication March 14, 2018

Table 1. Potential factors affecting data quality in clinical trials

Step in Clinical Trial Evolution	Potential Sources of Error/'Bad Data'
Instrument Selection	• Instruments measure inappropriate construct of interest
	• Psychometric characteristics of instrument are inappropriate to the study population
Rater selection (pre-qualification standards)	• Inadequate survey and verification of raters' experience, including:
	o knowledge of psychometric theory
	o knowledge of factors influencing subject performance
	o experience with the specific test instruments
	o experience with the study population
Rater training (methods of teaching study-specific protocol)	• Ambiguous administration or scoring instructions described in training materials including manuals
	• Failure to define appropriate and inappropriate prompts, cues, and corrections for commonly encountered participant response deviations
	• Failure to address subject response variations (regionalisms, accents, word derivatives, etc.)
	• Failure to address variability in instrument administration/scoring instruction between different studies and clinical use
Rater certification (standards of evaluating rater readiness to collect study-specific data)	• Inadequate evaluation of rater knowledge and skill in test administration (e.g. multiple choice questionnaires for decision of rater certification)
	• Minimal involvement of instrument content experts in preparatory and certification procedures
Test Administration	• Scoring worksheets lacking core administration instructions or important details of administration (prompts, start/stop rules, delay intervals)
Instrument Scoring	• Subjectivity/ambiguity/complexity of scoring criteria (e.g. paragraph recall, test of word meaning)
Data monitoring	• Source document review alone in the absence of ongoing rater performance evaluation with feedback (e.g. audio/video review)
	• Inadequate performance review of those who monitor study data (content experts and site monitors)
	Lengthy time interval between exam date and date of review
Utilization of 'expert' resources	• Difficult access to instrument experts for ongoing consultation and guidance in query resolution
	• Limited initiation of communication with experts (lack of time on part of the rater, lack of PI/site encouragement, etc.)
	• Motivational factors (perspective help is not needed, minimal consequences for errors, etc.)

possibility of conflicting training between ongoing trials. The natural evolution of rater experience from novice to expert - with refinement in technique and understanding of cognitive measures and participants being evaluated - may contribute to intra-rater performance variability over the full course of a study. Instrument administration manuals may be published before sufficient data is accumulated detailing administration issues that may arise in a given study population and they often do not provide the necessary detail for strategies of managing the range of subject responses and behaviors that may occur in a clinical population (see respective instrument manuals). Further, the available reliability measures from test development, when gathered, usually place emphasis on intra-rater test-retest reliability with highly qualified raters. Description of rater error in relation to specific administration and scoring guidelines are frequently minimal or absent. Taken together, reliability studies done during test development may not accurately predict the rater errors found in a multi-site clinical trial.

Unfortunately, there is little published literature demonstrating that the typical training raters receive during investigator meetings (whether done by sponsors or specialized training groups) or in online training modules is sufficient to yield competent rater performance in administration and scoring of the instruments. There is also little empirical evidence that widely used standard procedures for assessing rater qualification accurately reflect rater readiness for quality administration once the trial begins. Standards for rater certification in the use of a given instrument can be quite variable across studies, ranging from obtaining a passing score (e.g. 80%) on a multiple choice test where access to training materials is permitted during the examination

to more rigorous 'observation' of rater performance with a given instrument. While clinical trial sponsors often employ a specialized training group to manage these issues in a standardized way, it remains unclear which management strategy achieves the highest level of rater preparedness, demonstrated skill, and accountability to protocol among all who collect study data. Overall, there is little empirical evidence in the published literature that speaks to the evaluation of inter-rater reliability and/or error assessment across multiple sites engaged in a single clinical trial. Along with multiple other factors in trial design, either systematic (e.g. consistent over or under-scoring) or random variability in test results can result in inclusion of inappropriate subjects, exclusion of appropriate subjects, or affect detection of true clinical change (either false positive or more likely false negative), therefore error reduction is critical (5).

One way to address instrument and rater related factors impacting data quality has been to review inclusion/exclusion and outcome measure source documents during clinical trials. While it is routine for site monitors to review source documents for completeness of data, review of accuracy is often limited to verification of score calculation and of concordance between values entered on source documents and those entered in the larger database. Some studies involving cognitive screening or outcome measures will include review of the source documents and data by content experts; however, there is little published literature on the findings of these 'in trial' reviews with regard to frequency and nature of errors discovered. The focus of this paper is to summarize results from central review of inclusion instruments in a double blind clinical study of a potential memory enhancing compound for Age Associated Memory Impairment (AAMI). The paper aims to characterize the type and frequency of errors made on some common screening tests, present preliminary findings that provide insight into what instrumental and rater factors may influence the collection of quality data and suggest some areas still in need of further investigation.

Age Associated Memory Impairment reflects the detrimental effects of aging on memory that are not due to an underlying disease process such as Alzheimer's disease. As such, the criteria are based on decline in memory performance relative to performance at an earlier age. While the exact criteria are somewhat variable (6), the most common defining criteria has been subjective complaints of memory decline, adequate intellectual function, performance 1 standard deviation or more below the mean for a person in the 24–35 age range on standardized memory tests while not meeting criteria for dementia (7).

Methods

Raters were pre-qualified, trained and certified according to standard procedures. Potential raters were sent a Rater Experience Survey (RES) to determine if they met the minimal qualifications to move to the training portion of the study. The survey included questions pertaining to educational background, years of experience working in clinical trials research and with the study population of interest, and experience with various cognitive measures/instruments. Raters who met the predetermined qualification levels were considered qualified for specific study-related training while those that did not were reviewed on an individual basis to determine if they might still be appropriate raters for the study. Raters were then trained either at an investigator meeting or by review of a web-based recording of the IM for those who were unable to attend. Instrument training consisted of didactic review of instrument administration and scoring procedures. Certification on each instrument was established by performance on a multiple choice test.

Screening instruments administered in this study included the Mini Mental State Exam (MMSE; a brief screen of general cognitive function), the Memory Assessment Clinic Questionnaire (MAC-Q; a self-report questionnaire of perceived cognitive performance), the Geriatric Depression Scale (GDS; a self-report questionnaire of depressive symptoms), the Wechsler Memory Scale–Revised Verbal Paired Associate test, immediate recall condition (WMS-R VerPA I; a test of recall for word pairs), the Wechsler Memory Scale–Revised Logical Memory immediate and delayed recall condition (WMS-R LMI & LMII; a test of recall for two short stories), the Matrix Reasoning subtest of the Wechsler Abbreviated Scale of Intelligence (WASi; a test of non-verbal intelligence/reasoning ability), and the Vocabulary subtest of the Wechsler Abbreviated Scale of Intelligence (WASi; a test of verbal intelligence/knowledge of word meaning).

After test administration, clinical sites transmitted copies of test source documents by facsimile or email to the contract research organization (CRO). The CRO then posted the data on a secured portal which was reviewed by a neuropsychologist familiar with the instruments and specifics of administration, documentation and scoring required for this trial. Whenever questions arose in the review process or if the primary reviewer was unavailable, documents were reviewed by a second neuropsychologist. Reviewers communicated frequently to ensure they agreed on any determinations or decisions and on relevant discussion points for feedback calls to site raters. When administration/scoring issues were not clearly addressed in the test manual, the test publisher was contacted or an expanded set of scoring or procedural rules used by recognized groups (e.g. UDS, MOANS, etc.) was reviewed before a decision was finalized. All internal decisions were recorded to assure reviewer consistency throughout the study. Results of the review were entered into an Excel database for future analysis.

Main elements of review were detection of administration, procedural, documentation and scoring errors, including point assignments for responses. Since this level of review required access to participants' detailed responses, raters were required to document each subject's verbatim response on the source documents. For example, on the WMS-R Logical Memory test, raters were requested to document the subject's response verbatim and to indicate point assignment for each unit of information recalled correctly. Underlining exact matches to the target element on the score sheet was also permitted as long as raters also documented all non-exact language and provided some indication of the order of recall (e.g. by using arrows, etc.) so an exact replication of a participant's full response could be determined. This was necessary as scoring of some responses required proper context, some required a specific word or phrase, and some allowed a range of acceptable substitutions if the context of recall was supportive. Though review of source documents provides limited ability to detect administration or procedural errors that may have occurred during the test session, any documentation of response or scoring that suggested possible deviation from that described in the protocol were addressed with the rater by providing feedback through email or phone follow-up.

The results of each review were provided to the rater by email with specification of any corrective action to be taken. After initial review of the first 2 administrations of the test battery each rater was also contacted by phone and the results were discussed along with review of any suspected administration issues and fielding of questions from the rater. Throughout the rest of the study a rater was contacted selectively when they disagreed with the reviewer's findings, when observed to be making repetitive errors even after receiving feedback, or if a pattern of responses on source documents indicated the possibility of a significant administration or procedural deviation from protocol.

All instruments listed were monitored in their entirety with one exception. For the WASi Vocabulary test all responses were monitored for the first two administrations done by each rater. After feedback was given on the first 2 full administrations, seven item responses were randomly chosen and monitored for errors at each subsequent administration. In cases where the rater failed to show adequate improvement, further administrations were reviewed in their entirety. Only the first two administrations that included full reviews for all raters were included in the current report of WASi findings.

At the conclusion of the study, a sample of queries to which the rater agreed to make the recommended changes were cross-checked against the site source document to determine if changes were actually made. The electronic database (eDB) was then checked to determine if these changes were carried through to the eDB.

Results

Two hundred seventy-five complete screening evaluations administered by 28 raters (9.8+/-7.1 evaluations per rater; range 1 – 33) were reviewed for rater scoring error or deviation from standard test administration or other procedural guidelines. Data is presented as total errors and as average number of errors per administration (#/admin).

Wechsler Memory Scale–Revised: Logical Memory I and II

Source document review of the Wechsler Memory Scale–Revised Logical Memory immediate and delayed recall conditions (WMS-R LMI & LMII) revealed 772 instances (2.81/admin) of rater scoring error or deviation from administration/procedural guidelines. It should be noted that most scoring errors on this test can occur separately in both immediate and delayed recall conditions (e.g. errors in scoring participant's response), whereas some administration/procedural deviations can occur only once (e.g. improper delay interval between immediate and delayed recall).

The majority of detected procedural errors on WMS-R LM resulted from rater failure to provide sufficient documentation of the subject's response, as needed for central monitoring. Although raters were requested to record the full response verbatim and to indicate point assignment for each element recalled correctly, our review detected 125 instances (45%) where documentation of verbatim recall was incomplete. In 12 of these instances points were awarded for a correct response in the absence of verbatim documentation or underlining of scored story units to support scoring decisions. In the remaining instances of identified procedural error (113), scored elements were underlined but documentation of the verbatim or exact recall to support an assigned score was lacking. This allowed only monitoring the total calculation of underlined information units but not monitoring of the scoring decisions since context of recall could not be determined.

Test administration errors included 5 instances (0.02/admin) where the actual delay interval was longer than the maximum delay allowed and 1 instance (0.0036/admin) where it was shorter. There were 24 instances (0.087/admin) pertaining to incorrect calculation of the delay interval by a rater and 10 (0.036/admin) where the rater recorded the delayed recall time incorrectly on the worksheet so the true delay interval was in question. There were also 22 identified instances (0.08/admin) where the rater did not indicate if the story reminder had been given at the beginning of Logical Memory Delayed Recall which may have impacted the total score, as verbal elements in the reminder are not to be credited toward

the total recall score if they are re-stated in participant's subsequent recall.

Scoring errors included 50 instances (0.18/admin) where the rater miscounted points awarded on a given story line and 12 instances (0.044/admin) where the line score was correct but the total score was calculated incorrectly from the line scores. In 13 instances (0.047/admin) the rater did not document line totals or the total score on the source document and these could not be monitored.

Of particular concern in central review of the WMS-R LM test was that of the 150 administrations where verbatim documentation of story recall was available, there were 510 instances (3.40/admin) where the rater's scoring decision for an item response was in error. Of these, in 363 instances the rater failed to credit a correct response and in 147 instances the rater gave credit for an incorrect response.

Wechsler Memory Scale–Revised: Verbal Paired Associates I

Review of the Wechsler Memory Scale–Revised Verbal Paired Associates test (WMS-R VerPA I), immediate recall condition revealed 102 instances (0.37/admin) of rater scoring error or deviation from administration/procedural guidelines. The majority of findings related to a documentation requirement somewhat unique to this study. In order to properly monitor the tests, raters were asked to both record the subject's verbatim response for each word pair as well as the point score for that item (0 or 1). In 74 instances (0.27/admin) the rater either recorded the point score or the response but not both so accuracy could not be monitored. For those where both response and item score were provided (n = 208), 14 items (0.067/admin) were scored in error (given 0 instead of 1 point or vice versa). For simple addition errors, in 12 instances (0.044/admin) the total score was calculated incorrectly, and in 2 instances (0.007/admin) the total score was not calculated at all.

Vocabulary subtest of the Wechsler Abbreviated Scale of Intelligence

The Vocabulary subtest of the Wechsler Abbreviated Scale of Intelligence (WASi) revealed 162 instances (3.52/admin) of scoring or administration/procedural errors in 46 test administrations. As noted previously, only the first two administrations of the test by each rater were reviewed in their entirety and for the remainder of the study a subset of seven items were reviewed in each protocol. The findings presented here are only from the complete reviews of the first two administrations by each rater and represent 46 total administrations from 23 raters. Reported findings may underestimate the actual error rate, as in this sample 37 individual item responses from 8 administrations were illegible and could not be reviewed.

The primary procedural error discovered for the WASi Vocabulary test data was 55 instances (1.20/admin) where the rater failed to query a response as directed in the manual. Queries are to be made when a subject's initial response is of a quality deserving less than the maximum two points but which might be improved with addition of more detail for clarification. This rate of procedural error due to query failure likely underestimates the true rate (see discussion). Finally, additional low frequency procedural errors included not following the reversal rule or discontinuing the test too early or too late (2, 1 and 1 instances respectively).

On the Vocabulary subtest, it was noted that 73 responses (1.59/admin) were scored one point higher than they should have been according to the sample responses and scoring guidelines provided in the manual (over-scored). Most of these (93%) were instances where the subject's response was of a quality which should have been queried (i.e. probed for clarification of response ambiguity) for possible score improvement rather than just being given the higher score. Conversely, there were 24 instances (0.52/admin) where the subject should have received 2 points but was only given one point (under-scored). There were also 6 instances where unadministered reversal items were not correctly added to the total score.

Wechsler Abbreviated Scale of Intelligence: Matrix Reasoning

Review of source documents for the Matrix Reasoning subtest of the WASi revealed 69 instances (0.25/admin) of rater scoring error or deviation from administration/procedural guidelines. Compared to the WMS-R Logical Memory and the WMS-R VerPA I tasks, there were minimal documentation requirements on the Matrix Reasoning test but when documentation was incomplete, it was difficult to ascertain if error was attributable to improper test administration or incomplete documentation. For example, the rater's failure in 15 instances (0.054/admin) to circle a response or to indicate a score on the sample items could have resulted from either a documentation error or from failure to administer the sample items at all. Clear administration errors included 9 instances (0.033/admin) in which the reversal items 1-4 were administered when they should not have been, and 2 instances (0.0073/admin) when they were not administered but should have been. There were 3 instances (0.011/admin) where the wrong age-based start point was used, 2 (0.0073/admin) where the wrong age-based discontinue point was used, and 6 (0.022/admin) where the wrong discontinue point (based on number of sequential errors) was selected. Of greater concern were 15 instances (0.054/admin) where the sample items were included in the total calculation, resulting in an incorrect

total score. In contrast, there were 14 instances (0.051/admin) where raters failed to credit the 4 unadministered 'reversal' items as specified in scoring guidelines, again affecting the total score. Actual errors of addition affecting the total score, the kind of errors most often looked for in basic monitoring review, were a relatively rare finding in our central review (3 instances, 0.011/admin).

Mini Mental State Exam

Source document review of the Mini Mental State Exam (MMSE) interviews revealed 81 instances (0.29/admin) of rater scoring error or deviation from administration/procedural guidelines. Most were minor documentation issues such as failing to record age, level of consciousness, or mistakenly recording full name in the header (n = 40, 0.145/admin). Instances of deviation from standard administration procedure included giving the WORLD backwards task when serial 7's had already been administered (n = 6, 0.022/admin), not prompting properly when participant wrote part of an instruction as their sentence (n = 7, 0.025/admin), and accepting responses for orientation that were too general (n = 4, 0.145/admin). Instances of deviation from standard scoring included failure to adhere to scoring guidelines for serial 7's (n = 4, 0.0145/admin), the sentence (n = 10, 0.036/admin), or season (n = 2, 0.007/admin), not attending to details on the figure (n = 3, 0.011/admin) or repetition task (n = 2, 0.007/admin), and simple addition errors in calculating the total score (n = 3, 0.011/admin).

Memory Assessment Clinic Questionnaire

Review of the Memory Assessment Clinic Questionnaire (MAC-Q) source documents revealed 16 instances (0.058/admin) of rater scoring error or deviation from administration/procedural guidelines. The primary administration error identified was raters directly asking participants questions from the questionnaire rather than having them read and fill it out independently (n = 7, 0.025/admin). Scoring errors were also identified and mainly involved errors of addition (n = 7, 0.025/admin) and confusing the total score with the last item score (n = 2, 0.007/admin).

Geriatric Depression Scale

Source document review of The Geriatric Depression Scale (GDS) questionnaire revealed 14 errors (0.051/admin). The majority were simple addition errors (n = 13, 0.047/admin), with just one instance of failure to calculate the total score. Most commonly scores were in error by 1-2 points, with only one instance where the rater totaled the negative symptom score instead of the positive symptom score (e.g. scored '14' instead of '1').

Database change

As an exploratory investigation, during the site close-out visits a semi-random sample of queries from the screening instrument monitoring process were checked against the source documents to see if the recommended corrections were reflected in actual data changes made. Findings reported here are limited in that sampling was impacted by site monitor availability during the close-out visits. From a selection of 172 errors queried there were 27 instances (15.7%) where recommended changes were not made on both the source document and in the eDB. Of the 145 instances where changes were made on the source document, there were 13 instances (9.0%) where the corresponding change was not carried out in the eDB.

Discussion

Despite rater training which was consistent with current industry practice, ongoing monitoring by content experts of screening source documents detected multiple errors in some instruments. The tests that demonstrated the highest rate of discrepancy between site raters and centralized monitoring were the WMS-R Logical Memory test and the WASi Vocabulary test. While the WASi Vocabulary test was new to the majority of raters, most raters claimed experience with the Logical Memory test on the RES survey of prior experience. Errors on both of these tests appeared to be due to several factors, the most common being that raters did not consistently use the detailed scoring appendices provided and that variability in the requirement for contextual and/or conceptual support of recalled information likely contributed to confusion in application of scoring and prompting guidelines. For example, complexity in making scoring decisions for the WMS-R Logical Memory test arises because some responses receive credit only if information is situated in the proper recall context while for other responses context of recall is irrelevant. The Vocabulary test appeared challenging for raters due to the diversity of responses subjects provided when asked to "tell [me] what each word means". Point assignment (0-2) indicating quality of a given response relies on judgment of conceptual sophistication and in turn, of need for further probing of ambiguous responses. The frequency with which raters failed to properly query a subject in the Vocabulary test (55 instances in 46 administrations) likely underestimates the true rate since many subject responses were incorrectly scored at 2 points instead of 1 and therefore could not be queried for any further improvement. Had those items been correctly scored at the 1-point level it is likely more instances would have been found where the proper follow-up query was not given. These complexities for WMS-R Logical Memory and WASi Vocabulary tests introduced variability in administration and scoring not present for tests requiring simple recall of single stimulus points (e.g. words from

a word list), yes/no responses (e.g. a recognition task) or for tests utilizing a multiple choice response format (e.g. WASi Matrix Reasoning). Of note, while addition/summation errors represent a small percentage of the overall scoring errors found, they are the kind of errors most often assessed in standard site monitoring. This suggests that standard monitoring may miss a good number of errors that can be discovered on these instruments only with more in-depth review.

The error rate on the WMS-R Logical Memory test is troubling in that this has been and continues to be a test used extensively in memory and dementia research, both as a screening test and an outcome measure. It is a test with which the vast majority of raters in this study indicated having significant experience (on their RES) and while recall of stories is prone to responses that do not exactly fit the examples given in the test manual, we found many of the errors were made in scoring responses that directly matched examples given in the WMS-R manual appendix. This suggests raters were not properly using the scoring guide provided. It is noteworthy that during phone feedback discussion a few of the more experienced raters stated that they felt their scoring decision made more sense than what was in the manual and therefore did not want to adhere to the study standard.

Tests other than the WMS-R and WASi Vocabulary that were centrally monitored showed a lower error rate, likely due to the more straightforward scoring requirements. Most findings on these other tests reflected raters' failure to document responses completely in the unique way specified in the study protocol. Review of documentation on VerPA source documents that were complete revealed some errors that would not have been detected if only check marks had been required to indicate responses given (e.g. 14 incidents where the documented response was scored in error). While some of the errors appeared due to miscalculation, some appeared due to lack of familiarity with the test rules. For example, on the Matrix Reasoning task some raters showed difficulty in applying the reversal item rules and knowing when to credit unadministered items and not to credit sample items.

In addition to the main focus of the study (quantification of error rates on screening instruments), there were several incidental findings of interest, including those pertaining to rater qualification to administer the tests used in this study. Since training at investigator meetings often presumes some competency in basic psychometric theory, instrument and clinical experience, raters naïve in any of these areas are at greater risk of contributing to 'bad data' by making unintentional procedural errors of commission or omission. A statement of rater's prior experience with study instruments is valuable in estimating their current knowledge of as well as proficiency in test administration and in gauging training needs for the study of interest.

In this study and consistent with industry standard practice, potential raters were sent an RES to determine if they met the minimal qualifications to move to the training portion of the study. In-person follow-up verification of responses on the RES during the investigator meeting or through later phone contact revealed a number of surprising discrepancies between actual experience and that reported on the RES of some raters. While not quantified for this study, some raters who indicated they had used an instrument for many years and administered it more than 10 times in the past 12 months were discovered to have never administered it. Others had only read about it in an undergraduate class or observed someone else administering a similar test. When identified as having less than the minimum necessary experience such raters were not permitted to proceed with study-specific training. Though not formally examined in the current study, some common underlying reasons given for the discrepancies appeared to be that someone other than the candidate rater had completed the RES (e.g. the clinical coordinator), that the rater misunderstood which instrument(s) or time period of experience was being asked about on the RES, or that the rater was unaware or unclear about why that level of experience had been reported. Conversely but of equal importance, while raters with many years of testing and clinical experience are prized for their expertise, their increased comfort and reliance on prior experience with clinical instruments may in fact compromise strict adherence to study protocol as outlined in study-specific administration/scoring materials.

Another incidental finding and possible source of undiscovered errors came from the follow-up phone discussions with raters when feedback about performance was offered and questions answered. Since the phone contact was not done on a consistent basis but rather as needed based on expert evaluation of review findings, no formal quantitative analysis was done. Qualitatively it is worth noting, however, that phone discussion with several raters revealed a tendency to provide inappropriate prompts on the WASi Vocabulary (and to a lesser extent Matrix reasoning) that could affect psychometrics of the test. Some raters stated that when subjects were not succinct in their responses they might ask for a "briefer" response or ask to "just define the word". Though well-intentioned, such non-standard instructions to help the subject focus their answer may bias the natural verbal output being measured. When administration issues such as these were discovered in phone conversation, raters were advised about proper administration procedures and the reasoning underlying specific requirements.

Inconsistency of follow-through with data corrections, especially in the eDB, was another incidental finding and potential source of errors. The standard procedure for making data corrections was that if the rater agreed with a suggested correction they would change the item score

on the source document and make any resulting change to the instrument's total score. Then the rater or other site personnel would make the corresponding change in the eDB. In this study a subset of the queries was centrally checked to determine if changes agreed to by the rater were actually made on both the source document and in the eDB. We found instances of failure to make the changes on the source documents and in the database as well as some cases where the source document was changed but the database was not updated. Since the majority of queries were on the Logical Memory and Vocabulary tests for which there is some subjectivity in scoring of responses, it is possible that some initial agreements made by raters to change a score were later rescinded without notification. However, some cases were found where the source document was changed (indicating the rater accepted the change) but the eDB was not, so simple failure to follow through remains a strong possibility that should be monitored in future studies.

This study was a retrospective review of monitored data and is therefore limited by the original design of the monitoring protocol as well as the method of documenting findings, both of which were designed for practical utility during the trial but not optimized for later analysis. The method of data collection precluded standard statistical analysis. Error rates presented here may underestimate the error rate that would be found if raters were not given ongoing feedback throughout the study. We would anticipate that errors detected in this study of AAMI would share some similarities with those detected in different study populations; however, some differences in both the nature and extent of errors might also be expected in studies of clinical populations with greater cognitive impairment. Perhaps the greatest limitation of this study is that direct 'observation' of rater test administration, something not traditionally done in clinical trials and also not done in this study, limit findings of error to those discovered on source documents alone. Important deviations in test administration not evident on source documents such as incomplete or incorrect documentation of responses, inappropriate prompting or rewording of standardized instructions, and improper handling of stimulus materials would have remained undetected, contributing further to underestimation of scoring and procedural deviations.

New approaches to monitoring, particularly audio or audiovisual recording of the assessment ,offers enhanced opportunities for quality review. For example, in this study 45% of the logical memory tests were deemed to have insufficient verbatim documentation for full review, something easily overcome with audio recordings. Further, administration errors which are difficult to detect with source document review involving inappropriate instructions, cueing/prompting errors, source document recording errors, etc., could be monitored by audio review. This process also opens the opportunity for centralized scoring of the tests, minimizing inter-rater variability by using a few highly trained raters at a central site.

Technologies involving computer- or tablet-based assessment could also offer significant reduction in errors, particularly in the area of standardized administration, scoring, data capture and data transfer (8, 9). While our study indicated addition and data entry errors were relatively rare, there was an indication that carrying corrections through to the eDB may be problematic. However, without proper monitoring even device-based tests may be prone to many of the same errors found with pen and paper tests (10). For example, if a subject is very quickly recalling a list of words – tasking the rater to keep up - it would seem to matter little if the rater was recording the list on a piece of paper or on a tablet. While computer-based testing may have unique strengths in capturing measures of reaction time, fine motor skills and visiospatial perception, accuracy in measures involving rater judgment (story recall, vocabulary, etc.) may be susceptible to the same issues as pen and paper tests.

Findings reported here suggest several areas that would benefit from further investigation including verification of rater pre-qualifications, increased training focus on issues and common errors found with specific instruments/populations, performance-based evaluation for certification (such as live, scripted testing), enhanced central monitoring by content experts utilizing audio/video recording, ongoing feedback and recertification of raters, and central scoring of the more subjective tests. Follow-through with corrections to the data base also appears to be an area that needs further research. With the overall aim of reducing errors in data collected, comparison of various methods of rater training and performance evaluation for their effect on error reduction would be invaluable for future clinical trials research.

Exploration in all these areas has merit for guiding design and execution of future clinical research. Maximizing accuracy in measurement of cognitive function while minimizing burden to clinical trial sites and their staff is a goal that benefits all. Since there is undoubtedly a wealth of unpublished data from monitored clinical trials - both ongoing and completed - that could shed light on the various issues awaiting clarification, the authors encourage sponsors and monitoring groups to engage in retrospective analysis of their data, prospectively design future programs to more clearly address these questions, and to bring findings bearing on quality review procedures to publication adjacent to those describing clinical trial outcome findings.

Funding: None

Acknowledgements: Authors gratefully acknowledge the cooperation of Dart Neuroscience LLC in sharing the data for this analysis.

References

1. Hartling L, Hamm M, Milne A, Vandermeer B, et al. Validity and inter-rater reliability testing of quality assessment instruments. AHRQ Publication No. 12-EHC039-EF 2012. www.effectivehealthcare.ahrq.gov/reports/final.cfm. Accessed 6 April 2018
2. Schafer K, De Santi S, Schneider LS. Errors in ADAS-Cog Administration and Scoring May Undermine Clinical Trials Results. Curr Alz Res 2010;6:S496 – S497.
3. Connor DJ, Sabbagh MN. Administration and Scoring Variance on the ADAS-Cog. J Alzheimers Dis 2008;15:461–464.
4. Connor DJ, Sabbagh MN, Cummings JL. Comment on administration and scoring of the Neuropsychiatric Inventory (NPI) in clinical trials. Alz Dem 2008;4:390–394.
5. Petersen RC, Thomas RG, Aisen PS, et al. Randomized controlled trials in mild cognitive impairment: Sources of variability. Neurology 2017;88:1751-1758.
6. Hanninen T, Soininen H. Age-assoicated memory impairment. Normal ageing or warning of dementia? Drugs Aging 1997;11:480-9.
7. Crook, T, Bartus RT, Ferris SH, et al. Age associated memory impairment: Proposed diagnostic criteria and measures of clinical change - Report of a National Institute of Mental Health work group. Dev Neuropsych 1986;2:261-276.
8. Cummings J, Gould H, Zhong, K. Advances in designs of Alzheimer's disease clinical trials. Am J Neurodegener Dis 2012;1:205-216.
9. Zygouris S, Tsolaki M. Computerized cognitive testing for older adults: a review. Am J Alzheimers Dis Other Demen 2015;30:13-28.
10. Gates NJ, Kochan NA. Computerized and on-line neuropsychological testing for late-life cognition and neurocognitive disorders: are we there yet? Curr Opin Psychiatry 2015;28:165-172.