Examiner accuracy in cognitive testing in multisite brain-tumor clinical trials: an analysis from the Alliance for Clinical Trials in Oncology

Jane H. Cerhan, S. Keith Anderson, Alissa M. Butts, Alyx B. Porter, Kurt Jaeckle, Evanthia Galanis, and Paul D. Brown

Department of Psychiatry and Psychology, Mayo Clinic, Rochester, MN (J.H.C., A.M.B.); Alliance Statistics and Data Center, Mayo Clinic, Rochester, MN (S.K.A.); Department of Neurology, Mayo Clinic in Arizona, Scottsdale (A.B.P.); Departments of Neurology and Oncology, Mayo Clinic, Jacksonville, FL (K.J.); Department of Oncology, Mayo Clinic, Rochester, MN (E.G.); Department of Radiation Oncology, Mayo Clinic, Rochester, MN (P.D.B.)

Corresponding Author: Jane H. Cerhan, PhD, ABPP, Clinical Neuropsychologist, Division of Neurocognitive Disorders, Department of Psychiatry and Psychology, Mayo Clinic, 200 First St SW, Rochester, MN 55905, USA (cerhan.jane@mayo.edu).

Abstract

Background. Cognitive function is an important outcome in brain-tumor clinical trials. Cognitive examiners are often needed across multiple sites, many of whom have no prior testing experience. To ensure quality, we looked at examiner errors in administering a commonly used cognitive test battery, determined whether the errors were correctable upon central review, and considered whether the same errors would be detected using onsite electronic data entry.

Methods. We looked at 500 cognitive exams administered for brain-tumor trials led by the Alliance for Clinical Trials in Oncology (Alliance). Of 2277 tests examined, 32 noncorrectable errors were detected with routine central review (1.4% of tests administered), and thus removed from the database of the respective trial. The invalidation rate for each test was 0.8% for each part of the Hopkins Verbal Learning Test-Revised, 0.8% for Controlled Oral Word Association, 1.8% for Trail Making Test-A and 2.6% for Trail Making Test-B. It was estimated that, with onsite data entry and no central review, 4.9% of the tests entered would have uncorrected errors and 1.3% of entered tests would be frankly invalid but not removed.

Conclusions. Cognitive test results are useful and robust outcome measures for brain-tumor clinical trials. Error rates are extremely low, and almost all are correctable with central review of scoring, which is easy to accomplish. We caution that many errors could be missed if onsite electronic entry is utilized instead of central review, and it would be important to mitigate the risk of invalid scores being entered.

ClinicalTrials.gov identifiers: NCT01781468 (Alliance A221101), NCT01372774 (NCCTG N107C), NCT00731731 (NCCTG N0874), and NCT00887146 (NCCTG N0577).

Keywords

clinical trials | cognitive testing | neurocognitive

There is increasing interest in cognitive function as an outcome measure in brain-tumor clinical trials. Instruments used to measure cognition must be carefully selected to include requisite reliability, representative normative data, and population-specific validity. A 3-test cognitive battery has been recommended for this purpose^{1,2} that includes

the Trail Making Test Parts A and B (TMT-A and TMT-B),³ the Hopkins Verbal Learning Test-Revised (HVLT-R),⁴ and Controlled Oral Word Association (COWA).⁵ This battery has been used in multiple completed and ongoing clinical trials.^{6,7} The tests have well-established psychometric properties and clinical utility,⁸ and use of the battery across

studies allows direct comparison and potential pooling of data.

To provide testing for large trials, psychometric examiners are needed across numerous sites. Often these examiners have no prior experience in cognitive testing, so thorough training and certification is conducted to ensure valid and reliable results. Our objective in this study was to examine the accuracy of administration of these tests by certified clinical trial examiners. We looked at error rates in cognitive test administration using the aforementioned 3 tests, determined common error types, and determined whether they were correctable or if they invalidated the test. As a secondary aim, we considered whether errors would likely be detected if sites were conducting automated self-entry of results. The purpose was to inform quality-improvement measures for existing procedures, and to assess whether onsite entry would be prudent.

Materials and Methods

We looked at 500 cognitive exams administered for 3 brain-tumor trials led by the North Central Cancer Treatment Group (N107C, N0874, and N0577) and 1 trial led by the Alliance for Clinical Trials in Oncology (Alliance A221101). The North Central Cancer Treatment Group is now part of Alliance. N107C was a phase III trial comparing stereotactic radiosurgery with whole-brain radiation therapy in treating patients with brain metastases that have been removed by surgery. Neurocognitive progression was a primary outcome in this study. N0874 was a phase I/II study that compared vorinostat, temozolomide, and radiation therapy in treating patients with newly diagnosed glioblastoma multiforme. N0577 is actively recruiting and compares radiation therapy with concomitant and adjuvant temozolomide vs radiation therapy with adjuvant procarbazine, lomustine, and vincristine (PCV) chemotherapy in patients with grade II or grade III oligodendroglioma. A221101 is a phase III randomized, double-blind, placebo-controlled study examining armodafinil's ability to reduce cancer-related fatigue in patients with high-grade glioma.

Each participant whose data were used in this study signed an institutional review board-approved, protocol-specific informed consent document in accordance with federal and institutional guidelines. Data collection and statistical analyses were conducted by the Alliance Statistics and Data Center. The TMT-A, TMT-B, and COWA were administered in all 500 exams. The HVLTR was administered in about half of the exams (259) because of protocol differences. All exams had undergone central review (ie, rescoring by one Alliance reviewer [JHC]). We calculated the frequency of errors detected upon review and characterized the error types. Errors were coded as to whether they were corrected via central review vs invalidated the test.

Errors were also coded as to whether they would likely be detectable with onsite electronic data entry, or instead would require central review to be detected. Electronic onsite entry was presumed to involve basic score/data entry, and was not based on any specific data program or commercial product. Errors considered detectable and correctable using electronic entry included scenarios such

as leaving a field blank or entering implausible data into the field, both of which could trigger an electronic prompt. Some errors would be detected, but still not correctable, such as failure to time a timed test (examiner would realize this error when prompted to enter "total time tested" into the database). Other errors would not be detected by onsite entry and inaccurate scores would be entered (eg, proper noun accepted on COWA) or invalid tests accepted as valid (eg, administration technique failure on the TMT).

Training

Neurocognitive examiners for Alliance clinical trials must be certified by the designated neuropsychologist. To be certified, examiners must watch a training video, review a detailed administration-procedures document, complete a quiz, and submit a practice test. For the practice test, examiners must administer the battery to a colleague as a simulated patient. Some of the examiners in this study were trained by a partner group using the same methods, and were certified by reciprocity. Most examiners are research assistants, site coordinators, or nurses. A smaller number are physicians or other study personnel. Examiners are not qualified, by virtue of this training alone, to give any other tests, supervise others in testing, give these tests for other purposes, interpret test results, or share test results with examinees.

Tests

Hopkins Verbal Learning Test-Revised

The HVLTR is a word-list learning and recall memory test. It includes 3 components.

Immediate Recall (HVLTR IMM)

Three-trial learning of a 12-word list. Words are presented at a rate of 1 word every 2 seconds and the examinee is asked to recall after each trial. The score is the sum total number of words recalled on the 3 trials.

Delayed Recall (HVLTR DR)

A 20-minute delayed recall of the word list. The score is the total number of words recalled.

Delayed Recognition (HVLTR REC)

After delayed recall, distinguishing list words from foils in a 24-item yes/no format. Scoring requires subtracting the number of nonlist words incorrectly identified from the number of list words correctly identified.

Controlled Oral Word Association

COWA is a phonemic verbal fluency task in which the examinee must generate words beginning with a specified letter in each of 3 one-minute trials. The score is the sum total number of words generated across the 3 trials. Credit

is not given for proper nouns or for additional forms of the same word (eg, run, running).

Trail Making Test

TMT requires the examinee to rapidly connect numbered and/or lettered dots with a pen. For each part, a sample is administered to familiarize the examinee with the task. If the examinee makes a sequencing error, the examiner must keep timing, point out that an error was made, and direct the examinee back to the previous correct circle and instruct him or her to proceed in the proper sequence. The score for each part is the total time in whole seconds to complete the array.

Trail Making Test Part A

TMT-A requires the examinee to connect numbered dots in order. Time limit is 3 minutes.

Trail Making Test Part B

TMT-B requires the examinee to rapidly connect numbered and lettered dots with a pen, shifting back and forth between numbers and letters within sequence. Time limit is 5 minutes.

Results

In 2277 tests administered, we found 213 scoring or administration errors. Most of these were correctable upon central review. Thirty-two errors were not correctable, which is an invalidation rate of 1.4% of tests administered, with a rate of 0.8% for each part of the HVLT-R, 0.8% for COWA, 1.8% for TMT-A, and 2.6% for TMT-B (Table 1).

Hopkins Verbal Learning Test-Revised

In the trials that use this test, examiners submit test forms with scores for the HVLT-R for central review. Most errors involve incorrect or absent scoring, and are easily correctable because the actual patient responses are included. There were 20 errors on HVLT-R IMM (8%), 19 errors on HVLT-R DR (7%), and 30 errors on HVLT-R REC (30%). Two errors on each part (0.8%) were not correctable upon central review as they were administration errors (ie, due to the examiner translating to a language not approved by the protocol).

Error types included

Correctable Errors

- 52—didn't score.
- 11—incorrect score (all but one of these were on the REC subtest).

Noncorrectable Errors

- 06—protocol deviation.

Controlled Oral Word Association

The full test form is submitted for central review, including all of the words the participant produced. Of 500 administrations of COWA, we detected 108 total errors (22%), but only 4 (0.8%) had errors that were not correctable upon central review. Most errors on this test were scoring errors, correctable upon central review.

Error types included

Correctable Errors

- 24—test not scored.
 - 80—incorrect score. These were errors due to incorrect application of the scoring rules. Twenty administrations had more than one mistake, indicating a small subset of examiners were less attentive to the scoring rules.
- Subtypes of these scoring violations included the following:
- 43—accepted proper noun (eg, "Peterbilt").
 - 30—accepted different form of same word (eg, run, running).
 - 12—other scoring error (eg, accepted "locomotion" but not "locomotive," accepted "afford" and "filander" for the letter "F").
 - 16—accepted same word repeated.
 - 8—accepted nonword (eg, "litch," "fixty," and "whopsided").

Noncorrectable Errors

- 2—did not time the test.
- 2—gave the test in a language not allowed in the particular protocol.

Trail Making Test Part A

The raw test forms are submitted for central review, along with a score page. There were 15 errors on TMT-A. Of 500 administrations of TMT-A, we detected 15 errors (3%), with 9 (1.8%) having errors that were not correctable upon central review.

Error types included

Correctable Errors

- 6—recording errors. Most of these included leaving portions of the scoring form blank but recording the time on the test itself.

Table 1 Rate of Noncorrectable Errors

Test	HVLT-R IMM	HVLT-R DEL	HVLT-R REC	COWA	TMT-A	TMT-B	Total
Number of tests given	259	259	259	500	500	500	2277
Invalidating errors (% of tests given)	2 (0.8%)	2 (0.8%)	2 (0.8%)	4 (0.8%)	9 (1.8%)	13 (2.6%)	32 invalidated subtests (1.4% of tests administered)

Abbreviations: COWA, Controlled Oral Word Association; HVLT-R DEL, Hopkins Verbal Learning Test-Revised Delayed Recall; HVLT-R IMM, Hopkins Verbal Learning Test-Revised Immediate; HVLT-R REC, Hopkins Verbal Learning Test-Revised Recognition; TMT-A, Trail Making Test Part A; TMT-B, Trail Making Test Part B.

Noncorrectable Errors

9—administration error: This means there is evidence the test was administered incorrectly, such as not pointing out errors in real time or not timing the test.

Trail Making Test Part B

Of 500 administrations of TMT-B, we detected 21 errors (4%), with 13 (2.6%) having errors that were not correctable upon central review.

Error types included

Correctable Errors

8—recording errors. Described above in TMT-A.

Noncorrectable Errors

13—administration errors: Described above in TMT-A. Two of these were implausible entries (eg, “3 seconds” to complete a trial).

Onsite Electronic Entry

Hopkins Verbal Learning Test-Revised

We determined that 52 of the 69 errors detected with central review would also be detected with a basic onsite electronic score-entry system. The remaining 17 errors would not be detected and incorrect data could be entered. However, if the system was programmed to do the scoring, which would be relatively easy for this test, only 2 errors would go undetected and uncorrected (protocol deviations), potentially allowing incorrect data into the database for those 2 cases. In contrast, these would be detected and invalidated with central review, and as such would be (correctly) left out of the database.

Controlled Oral Word Association

On COWA, we determined that 24 of the 108 errors detected with central review would also be detected and corrected with a basic onsite electronic score-entry system as they were due to failure to score, and examiners could be cued to complete all fields. The remaining 84 of the 108 errors would not have been detected with electronic onsite entry and inaccurate scores could be entered. Of these, 80 were simple scoring-rule errors resulting in potential entry of a score that would be off by only 1 or 2 points. This issue would be difficult to avoid with electronic entry as it would be very difficult to program an onsite entry system to apply the scoring rules for this test. The other 4 undetected errors were due to invalid procedures and these scores could be entered and not detected with onsite entry, whereas they were invalidated by central review.

Trail Making Tests Part A and Part B

On TMT-A, of the 15 errors detected with central review, the 6 recording errors (did not enter time on score sheet) would also be detected by onsite entry and examiners cued to correct. The 9 errors due to departures from standardized procedure would not be detected, potentially allowing invalid data into the system. Similarly, on TMT-B, of 21 errors, the 8 recording errors would be detected and corrected, whereas the 13 administration failure errors would likely go undetected with the possible exception of the 2 errors due to implausible entries. These 2 implausible entries might have been detected and properly invalidated by onsite entry.

Discussion

We examined 500 neurocognitive batteries administered for brain-tumor clinical trials utilizing recommended cognitive tests (TMT, HVLT-R, and COWA). Of 2277 tests administered, we found 213 scoring or administration errors. Most of the errors were minor and corrected upon central review, with uncorrectable errors in just 32 (1.4%) of the tests administered. Central review is easy to

accomplish, is not time consuming, and can be conducted by a trained technician. Additional costs are low, as test materials can be examined directly or via electronic copies.

The low rate of invalidating errors is encouraging, and supports the robust nature of cognitive testing outcomes and the continuation of centralized review for the studies Alliance supports. In a smaller study, Regine et al⁹ looked at examiner errors in 130 clinical trial administrations of essentially the same battery examined in our study. They used an earlier version of the HVLIT without delayed recall, and they also examined an attention test not used here. They looked for 2 types of errors: errors in administering the test, and errors in scoring, reporting total number of “unusable” administrations. Using these criteria, they found that 10% of 1300 individual test administrations were “unusable” because of error. This is similar to our finding that 213 of 2277 tests had errors (9.4%), but we had a much lower rate of “unusable” tests (1.4%) after making corrections through central scoring.

The overall number of uncorrectable TMT-B tests was low (2.6%) but still slightly higher than the other tasks because of procedural deviations. This test can be more complicated to administer as examiners have to watch closely and correct participant sequencing errors in real time. To address this, we added the requirement that practice participants make errors during training, resulting in a marked decrease in invalid tests. We also observe that more errors occur when the patient is more impaired, although this was not empirically examined in the current study. Other enhancements of training for this test could include periodic refreshers, such as emailing memos or videos demonstrating common errors, throughout the duration of the study.

We advise caution around using these error-rate results to inform test selection for clinical trials. It might seem appealing to choose the tests with the lowest error rate, but given very low error rates overall, there are other important considerations including the test’s utility in detecting impairment and change in the population of interest, and in providing balance to the battery so that multiple cognitive domains are covered. A somewhat more complex task, such as the TMT-B, while somewhat more difficult to administer, might detect important aspects of executive functions or tap into changes a simpler task might miss.

We estimated that, of the 213 errors detected by central review, 112 (52.6%) would go undetected if onsite data entry was used alone, leading to an undetected error rate of 4.9%. Most of these errors were scoring rule violations on the COWA (such as giving credit for ineligible words). It would be cost prohibitive for most trials to develop an onsite scoring system that could apply all of the scoring rules to COWA words. That said, the typical COWA scoring error affects the total score by only one point, which arguably has minimal overall impact. A more serious issue is that 30 of the 32 invalid administrations (1.3% of all tests given) would most likely not have been detected using onsite data entry alone. This means invalid results could be entered into the database in those cases, whereas central review allows these to be detected and removed. Furthermore, when procedural errors are detected centrally, we follow up by contacting the examiners with remedial instruction thus preventing future loss of data.

Our intent is not to rule out onsite entry, but to consider important features that could lead to undetectable errors and optimize quality control. Some computerized tests could be appealing for this purpose, and this will likely be the norm at some point in the future. However, currently most lag well behind traditional tests in terms of necessary accumulated validity evidence and clinical interpretability.¹⁰ They also often focus disproportionately on processing speed, when measures of higher-order executive functions (eg, TMT-B) have proven important as outcome measures in previous brain-tumor trials.⁶

In summary, cognitive test results are important and robust outcome measures for brain-tumor clinical trials. After central review of administration and scoring, we found error rates were extremely low for the specific tests used (COWA, TMT, and HVLIT-R). It should be noted that these low error rates apply only when training is meticulous and monitored by qualified personnel. When performed well, the process is cost efficient (especially when compared with more expensive outcomes such as imaging), well received by study sites, and often there is training reciprocity from other studies using the same battery. Central review is easy to accomplish, and can both detect errors and prevent future errors by notifying examiners of any deviations. We caution that many errors detected with central review could be missed with onsite electronic entry. We provide some caution around using simpler tests or computerized test products that have not been proven to provide equivalent breadth and interpretability.

Funding

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under award number UG1CA189823 (Alliance for Clinical Trials in Oncology National Cancer Institute Community Oncology Research Program Grant) and U10CA180790. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Joly F, Giffard B, Rigal O, et al. Impact of cancer and its treatments on cognitive function: advances in research from the Paris International Cognition and Cancer Task Force Symposium and update since 2012. *J Pain Symptom Manage*. 2015;50(6):830–841.
2. Wefel JS, Vardy J, Ahles T, Schagen SB. International Cognition and Cancer Task Force recommendations to harmonise studies of cognitive function in patients with cancer. *Lancet Oncol*. 2011;12(7):703–708.
3. Reitan RM. *Trail Making Test: Manual for Administration and Scoring*. South Tucson, AZ: Reitan Neuropsychological Laboratory; 1992.
4. Brandt J. The Hopkins Verbal Learning Test: development of a new memory test with six equivalent forms. *Clin Neuropsychol*. 1991;5(2):125–142.
5. Benton AL, Hamsher K. *Multilingual Aphasia Examination. Manual of Instructions*. 2nd ed. Iowa City, IA: AJA Associates; 1989.

6. Brown PD, Jaeckle K, Ballman KV, et al. Effect of radiosurgery alone vs radiosurgery with whole brain radiation therapy on cognitive function in patients with 1 to 3 brain metastases: a randomized clinical trial. *JAMA*. 2016;316(4):401–409.
7. Gilbert MR, Dignam JJ, Armstrong TS, et al. A randomized trial of bevacizumab for newly diagnosed glioblastoma. *N Engl J Med*. 2014;370(8):699–708.
8. Strauss E, Sherman EMS, Spreen O. *A Compendium of Neuropsychological Tests*. 3rd ed. Oxford, England, UK: Oxford University Press; 2006.
9. Regine WF, Schmitt FA, Scott CB, et al. Feasibility of neurocognitive outcome evaluations in patients with brain metastases in a multi-institutional cooperative group setting: results of Radiation Therapy Oncology Group trial BR-0018. *Int J Radiat Oncol Biol Phys*. 2004;58(5):1346–1352.
10. Bauer RM, Iverson GL, Cernich AN, Binder LM, Ruff RM, Naugle RL. Computerized neuropsychological assessment devices: joint position paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. *Arch Clin Neuropsychol*. 2012;27(3):362–373.