Digital cognitive testing using a tablet-based app in patients with brain tumors: a single-center feasibility study comparing the app to the gold standard

Vicki M. Butenschoen, MD, MSc, Jasmin Ahlfeld, MD, Bernhard Meyer, MD, and Sandro M. Krieg, MD, MBA

Department of Neurosurgery, Technical University of Munich, School of Medicine, Klinikum rechts der Isar, Munich, Germany

OBJECTIVE Healthcare digitization has led to increasing tablet-based apps to improve diagnostics, self-discipline, and well-being in patients. Moreover, patient-reported outcome measures are crucial for optimized treatment, with superior applicability if independent from patient visits. Whereas most uses cover health maintenance, only a few studies have focused on cognitive testing in neurosurgical patients despite its nature as one of the most integrative outcome measures in neurooncology.

METHODS The authors performed a prospective single-center feasibility study including neurosurgical patients affected by intraaxial tumors and healthy subjects, testing cognitive function by using a digitized app-based approach and conventional paper-and-pencil (PP) tests. Healthy subjects underwent follow-up testing for retest reliability.

RESULTS The authors included 24 patients with brain tumor and 10 healthy subjects, all of whom completed both tests. Equivalent mean performance results were found in the tablet-based digital app and PP counterparts; whereas the digital approach had shorter test duration in patients (29.9 minutes for PP vs 21.9 minutes for app, p = 0.019) and in the healthy cohort (23.2 minutes for PP vs 16.4 minutes for app, p = 0.003), patients with brain tumor scored lower when both test strategies were applied. Results were consistent in healthy subjects after a median of 3 months.

CONCLUSIONS Cognitive function assessment is feasible using a digitized tablet-based app, with equivalent results to those of PP tests in healthy subjects and patients with brain tumor. Thus, this approach allows much closer follow-up independent of patient visits and might provide a viable option to improve patient follow-ups.

https://thejns.org/doi/abs/10.3171/2022.3.FOCUS21726

KEYWORDS cognition testing; brain tumor; digital healthcare
MRI for all included patients with brain tumor. Patients experiencing a known dementia or severe aphasia were excluded. The recruitment period was September 2020 to March 2021.

**Study Design**

We conducted a prospective single-center cohort analysis as a pilot study and compared two distinct study populations: healthy subjects without known cognitive impairment, and patients affected by newly diagnosed brain cancer (either glioma or metastases). In the healthy cohort, the testing was performed twice to assess the reliability of both cognitive testing approaches. All test subjects underwent the PP and digital app-based cognitive testing for comparison. For both testing modalities, the subjects were supported by a professional interrogator, offering help in cases of difficulties encountered during the PP tests. No instruction help was given during the digital testing. The order in which the test versions were presented was counterbalanced (50% underwent the digitized test first, 50% completed the PP test first) to control for any bias, such as learning effects.

**Health App (cognition app)**

We used a tablet-based, user-friendly mobile app to test cognitive function in healthy subjects and patients. The app includes tests on memory (word list presentation with short-term and delayed recall after 5 minutes, analogous to the Hopkins Verbal Learning Test); vocabulary (semantic fluency and Controlled Oral Word Association Test from the Multilingual Aphasia Examination, modified from Benton et al.); and logic (trail-making test, sorting numbers for arithmetic fluency, and Symbol Digit Modalities Test), and provides automated scoring and standardized administration instructions.

**Statistical Analyses**

Statistical analyses were performed using the SPSS software version 26 (IBM Corp.). We compared categorical data using the chi-square or Fisher’s exact test as needed. The mean values were compared using the independent-samples t-test, and test reliability was evaluated using the paired-samples t-test. All tests were performed 2-sided, and a p value < 0.05 was considered significant.

**Ethics**

This study is in accordance with ethical standards outlined in the Declaration of Helsinki, and approval was obtained from our local ethics committee prior to subject recruitment. All patients and healthy subjects gave written informed consent.

**Results**

**Study Population**

In total, 10 healthy subjects and 24 patients affected by intraaxial tumors were screened and included for analysis between September 2020 and March 2021.

**Patients With Brain Tumors**

The mean age of the patients with tumors was 57 years (range 25–83 years). Female and male sex were equally represented (both n = 12). Regarding the medical history of our cohort population, 3/24 patients (12.5%) experienced a recurrent depression disorder and were receiving antidepressant administration, and 1 patient (4.2%) was affected by a mild aphasia. The median KPS score was 90%, significantly differing from the 100% score in the healthy population (p = 0.034). In total, 11 patients (45.8%) were affected by a left-sided lesion, 12 (50%) a right-sided lesion, and 1 bilateral glioblastoma. Twenty-two patients (91.7%) were right-handed and 2 were left-handed. Eighteen patients (75%) were affected by a WHO grade 2–4 astrocytoma, 3 (12.5%) an oligodendroglioma, 1 a central neurocytoma, and 1 a ganglioglioma. One patient had a recurrent glioblastoma (WHO grade 4).

**Healthy Cohort**

Ten healthy subjects were tested as a control population. The subjects were only included if they denied any pharmaceutical neurological treatment and had never undergone an operation for intraaxial tumors. The mean age was 47.1 years (range 25–85 years); 60% were female, and 40% were male. There was no significant difference between the patient and healthy population regarding age (p = 0.172) and sex (p = 0.595). None of the healthy subjects experienced psychological and neurological disorders, and past medical histories only included autoimmune diseases such as hypothyroidism (4/10, 40%) and diabetes type 2 (1/10, 10%; the patient with diabetes also had hypothyroidism).

The healthy cohort tended to be more educated than our patient cohort (median years of education 15 vs 12 years), but the difference remained statistically insignificant (p = 0.199) (Table 1).

**Test Pracitcality**

Most patients were able to participate in both tests—the written PP test and the app-based assessment.
Regarding time duration to finish the assessment, the PP tests required longer than the app-based approach (mean 27.9 vs 20.3 minutes, range 14–52 vs 11–54 minutes; p = 0.014). The difference in duration was observed in both populations: in our patient cohort (29.9 vs 21.9 minutes, p = 0.019) and in our healthy cohort (23.2 vs 16.4 minutes, p = 0.003) (Fig. 1).

After performing both tests, all subjects were asked to complete a form investigating the practicality (how intuitive are the instructions?), the time given for each task (did the subjects have enough time to fulfill the tasks?), the clarity of the instructions (no misunderstandings in performing the required tasks), and the subjective preference for the tests (which test would you recommend continuing with?). Answers were rated from 0 (not at all) to 10 (perfect agreement). Regarding their practicality, both tests were considered practical (mean score 8.4 vs 8.8, p = 0.463) and clear regarding the instructions (mean score 8.9 vs 8.6, p = 0.258). Significant differences in favor of the PP test were identified regarding the definition of the instructions (9.1 vs 8.7, p = 0.028) and the time given to finish the required tasks (9.3 vs 8.7, p = 0.050) (Table 2).

Difficulties encountered in the patient cohort during the digital evaluation (Fig. 2) included the following: touchscreen handling, especially in patients experiencing motor deficits (n = 1); short reaction times in healthy subjects (n = 1); and delayed reaction in patients with glioma (n = 3).

Test Validity

The app-based testing included standardized word
naming and memory testing similar to the Hopkins Assessment of Naming Actions (HANA). The applied tests furthermore included the following: word memory (12 words) assessed three times in a row and recalled at the end of the test battery, vocabulary testing, symbol digit recognition, and trail making. All included tests were designed along the same lines as the validated standard cognitive tests.

We found no significant difference between the app-based approach and PP tests in vocabulary and word memory performance tasks (first round p = 0.335, second round p = 0.0842, third round p = 0.173, and end recall p = 0.337). Score equivalence was observed in all categories assessed.

Comparing the patient and healthy cohort, patients with tumors scored significantly lower in all categories, including verbal memory, word naming tasks/verbal skills, and trail-making tasks (p < 0.001 to p = 0.014). All results are summarized in Table 3.

### Test Reliability

In our 10 healthy subjects we performed retesting after a median of 113 days (approximately 3 months; range 28–245 days). All healthy subjects were asked to undergo the same testing again and the results were compared with paired-samples t-testing. We found similar results after 3 months with shorter test durations (mean duration: app 14.6 vs PP 19.2 minutes, range 13–33 minutes; p = 0.002), without any significant differences regarding the different tasks performed (Table 4). In our healthy population, both testing approaches were therefore considered reliable.

### TABLE 2. Comparing duration, perceived practicality, timing, clarity of instructions, and definitions of the cognitive testing

<table>
<thead>
<tr>
<th></th>
<th>Duration (mins)</th>
<th>Practicality</th>
<th>Timing</th>
<th>Definition of Instructions</th>
<th>Clarity of Instructions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>App  PP</td>
<td>App PP</td>
<td>App  PP</td>
<td>App  PP</td>
<td>App  PP</td>
</tr>
<tr>
<td>Healthy subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>16.4 23.2</td>
<td>9.4 9.0</td>
<td>9.8 10.0</td>
<td>9.5 9.7</td>
<td>9.6 9.5</td>
</tr>
<tr>
<td>Median</td>
<td>15.0 20.5</td>
<td>10.0 9.0</td>
<td>10.0 10.0</td>
<td>10.0 10.0</td>
<td>10.0 10.0</td>
</tr>
<tr>
<td>Pts w/ tumors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>21.9 29.9</td>
<td>8.0 8.7</td>
<td>8.3 9.0</td>
<td>8.4 8.9</td>
<td>8.5 8.1</td>
</tr>
<tr>
<td>Median</td>
<td>19.0 28.5</td>
<td>8.0 10.0</td>
<td>10.0 10.0</td>
<td>10.0 10.0</td>
<td>10.0 8.5</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>20.3 27.9</td>
<td>8.4 8.8</td>
<td>8.7 9.3</td>
<td>8.7 9.1</td>
<td>8.9 8.6</td>
</tr>
<tr>
<td>Median</td>
<td>18.0 25.5</td>
<td>9.0 9.0</td>
<td>10.0 10.0</td>
<td>10.0 10.0</td>
<td>10.0 9.5</td>
</tr>
</tbody>
</table>

App = app-based test. Boldface type indicates statistically significant differences in duration and ratings (scores 0–10).
Patients undergoing adjuvant treatment after surgery were not included for follow-up testing.

**Discussion**

The digital app testing the cognition of healthy subjects and patients with glioma showed consistent results, with deficits in speech recognition and vocabulary testing. The digital app was time-efficient and feasible and showed a high internal validity with a low rate of performance issues. We found no significant performance differences in using the app-based approach compared with the PP tests. Patients scored significantly lower than healthy subjects, but scores did not depend on the assessment technique.

**Limitations**

There are some limitations to our study that need to be discussed. First, we only provide a small sample size of test subjects and we did not include patients with strong cognitive impairments who would indeed require personal instructions and personal help to complete the tests. Neither did we assess anxiety potentially causing cognitive deficits. Patients needing support could probably use both approaches, but would require personal help with the app-based approach due to its limited timing, inadequate tool flexibility, and touch pad handling. We did not investigate the comparison in patients with severe cognitive deficits; therefore, our results can only apply to patients with no or only mild cognitive impairment.

To evaluate the reproducibility of our test approaches, we only tested healthy subjects after a median of 3 months. We did not include the patient cohort due to their adjuvant treatment (mostly radiation therapy), which remains a strong limitation given that the potential benefit of follow-up examinations in impaired patients presents an important aspect in health digitization. Furthermore, we expect cognitive function to decline after surgery and at follow-up.12

The app-based approach was completed under continuous support and observation from one of the study physicians (J.A.). Results from the digital app were double-checked (written down manually and saved digitally) and showed a difference especially in the word recognition tasks. While silently pronounced or mispronounced words could still be recognized for the PP tests, the app was less forgiving and only saved words when pronounced and stated clearly. Digital-based cognitive testing should therefore be optimized to recognize words more broadly, especially in vocabulary and word memory tasks.

### TABLE 3. Results obtained in healthy subjects and our cohort of patients with tumors from word memory testing performed using the digitized cognitive app and PP tests

<table>
<thead>
<tr>
<th></th>
<th>Word Memory 1</th>
<th>Word Memory 2</th>
<th>Word Memory 3</th>
<th>Recall Words</th>
<th>Word Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>App</td>
<td>PP</td>
<td>App</td>
<td>PP</td>
<td>App</td>
</tr>
<tr>
<td>Healthy subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>74.2</td>
<td>64.4</td>
<td>87.5</td>
<td>91.7</td>
<td>97.5</td>
</tr>
<tr>
<td>Median</td>
<td>75.0</td>
<td>66.7</td>
<td>87.5</td>
<td>91.7</td>
<td>100.0</td>
</tr>
<tr>
<td>Pts w/ tumors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>45.5</td>
<td>44.4</td>
<td>63.2</td>
<td>62.2</td>
<td>70.1</td>
</tr>
<tr>
<td>Median</td>
<td>41.7</td>
<td>45.8</td>
<td>66.7</td>
<td>66.7</td>
<td>75.0</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>53.9</td>
<td>50.3</td>
<td>70.3</td>
<td>70.8</td>
<td>78.2</td>
</tr>
<tr>
<td>Median</td>
<td>50.0</td>
<td>50.0</td>
<td>75.0</td>
<td>83.3</td>
<td>87.5</td>
</tr>
</tbody>
</table>

Results (expressed in %) were equivalent in both testing modalities. Word memory 1–3 designates first, second, and third assessments.

### TABLE 4. Results obtained in healthy subjects 3 months after using the digitized cognitive app and PP tests

<table>
<thead>
<tr>
<th></th>
<th>Word Memory 1</th>
<th>Word Memory 2</th>
<th>Word Memory 3</th>
<th>Word Recognition</th>
<th>Recall Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>App</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>70.8</td>
<td>90.0</td>
<td>96.7</td>
<td>99.6</td>
<td>95.0</td>
</tr>
<tr>
<td>SD</td>
<td>14.8</td>
<td>9.5</td>
<td>5.8</td>
<td>1.3</td>
<td>5.8</td>
</tr>
<tr>
<td>PP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>93.3</td>
<td>97.5</td>
<td>95.8</td>
<td>48.7</td>
<td>95.8</td>
</tr>
<tr>
<td>SD</td>
<td>8.6</td>
<td>5.6</td>
<td>5.9</td>
<td>2.0</td>
<td>5.9</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>82.1</td>
<td>93.7</td>
<td>96.2</td>
<td>74.2</td>
<td>95.4</td>
</tr>
<tr>
<td>SD</td>
<td>16.5</td>
<td>8.5</td>
<td>5.7</td>
<td>26.1</td>
<td>5.7</td>
</tr>
</tbody>
</table>

Results (expressed in %) were equivalent in both testing modalities. Word memory 1–3 designates first, second, and third assessments.
Interrater Reliability

Whereas most PP tests require a medical professional instructing the patient, the digital app aimed to be self-explanatory without requiring help to follow the instructions. Given that the app relies solely on the patient him- or herself, the reliability may benefit from less rater-dependent instructions and interpretation of the results and a more automated scoring system. Rater effects and rater bias (referring to differences in test results depending on the examiner’s interpretation of the results) have commonly been described in psychological studies focusing on language and performance tasks. Using an automated digital, self-explanatory app may reduce rater inconsistency.

Digital Security and Privacy

Addressing health record digitization, the challenges of digital security and protection of sensitive cognitive data have to be met for tested subjects. Security measures and responsibilities should be defined prior to using new digital techniques, and test subjects should be well informed and aware of the different parties accessing their personal data. In our study, the app was used on a single device (a password-locked tablet), and healthcare data sent to a third party for analysis were anonymous, without a possibility to be traced back to any patients or healthy subjects. No other data were saved using the app—that health data such as diagnosis, nor age or sex. All test subjects were informed about the inherent risks and asked for permission and informed consent to acquire digital data. The anonymous data and results were obtained from the digital app but not saved in any electronic patient health record.

Diagnostic Tools Versus Health Apps

In our study, we did not aim to analyze potential benefits of the digital app; rather, we sought to present a non-inferiority of the digital alternative compared with the time- and paper-consuming PP tests. Whereas most available health apps seek to improve or monitor a medical condition and well-being, the app presented here is only used for diagnostic purposes and the testing is conducted by a responsible physician. The acquired data were not used to compare any treatment outcome, given that we performed this as a pilot feasibility study. Of course, many alternatives are available nowadays, such as the Brief Assessment of Cognition (BAC) app for schizophrenia17 and apps for telerehabilitation,18 to name only a few.

Potential Benefits

Many benefits may arise from digitizing health data, such as more efficient cognitive testing, when privacy and security of patient data can be guaranteed. Analyzing cognitive impairment due to drug interactions or neurosurgical treatment may be simplified, and data can be more easily compared and continuously captured when saved digitally. When applied routinely, digital apps may reduce the resources usually required to assess cognitive impairments, such as nurses or physicians. Because of the significantly shorter test duration, patients who are intolerant to longer test durations may benefit from digitized alternatives.

Although digitized cognitive testing has been described in healthy subjects and patients affected by dementia before, our study presents the first use of digitized cognitive testing in neurosurgically treated patients, comparing it to healthy subjects and showing its equivalence in both populations.

Conclusions

Using a digitized cognitive testing app in neurosurgically treated patients and healthy subjects, we found comparable results consistent with the traditional PP counterpart. Patients reported high satisfaction regarding the understandability and usability of the digital approach, while reporting difficulties with regard to the time given for each task and issues encountered in word recognition tasks. Thus, this approach allows much closer follow-up independent of patient visits and might provide a viable option to improve patient follow-ups.

References


Disclosures
Dr. Krieg is a consultant for Ulrich Medical, and received honoraria from Spineart Deutschland GmbH, Nexstim Plc, Medtronic, and Carl Zeiss Meditec. Drs. Butenschoen, Krieg, and Meyer received research grants from and are consultants for Brainlab AG. Dr. Meyer received honoraria, consulting fees, and research grants from Medtronic, Icotec AG, and Relievant Medsystemy, Inc.; honoraria and research grants from Ulrich Medical; honoraria and consulting fees from Spineart Deutschland GmbH and DePuy Synthes; and royalties from Spineart Deutschland GmbH.

Author Contributions
Conception and design: Butenschoen, Ahlfeld. Acquisition of data: Butenschoen, Ahlfeld. Analysis and interpretation of data: Butenschoen, Krieg. Drafting the article: Butenschoen. Critically revising the article: Krieg. Reviewed submitted version of manuscript on behalf of all authors: Butenschoen. Administrative/technical/material support: Meyer.

Correspondence
Vicki M. Butenschoen: Technische Universität München, Klinikum rechts der Isar, Munich, Germany. vicki.butenschoen@tum.de.