How impaired is too impaired? Exploring futile neuropsychological test patterns as a function of dementia severity and cognitive screening scores

Andrew M. Kiselica[*,1], **Ellen Johnson**[1,2], **Jared F. Benge**[3]

[1]Department of Health Psychology, University of Missouri, Columbia, Missouri, USA

[2]Department of Psychology, Ohio University, Athens, Ohio, USA

[3]Department of Neurology, University of Texas at Austin, Texas, USA

Abstract

Some older adults cannot meaningfully participate in the testing portion of a neuropsychological evaluation due to significant cognitive impairments. There are limited empirical data on this topic. Thus, the current study sought to provide an operational definition for a futile testing profile and examine cognitive severity status and cognitive screening scores as predictors of testing futility at both baseline and first follow-up evaluations. We analysed data from 9,263 older adults from the National Alzheimer's Coordinating Center Uniform Data Set. Futile testing profiles occurred rarely at baseline (7.40%). There was a strong relationship between cognitive severity status and the prevalence of futile testing profiles, $\chi^2(4) = 3559.77$, $p < .001$. Over 90% of individuals with severe dementia were unable to participate meaningfully in testing. Severity range on the Montreal Cognitive Assessment (MoCA) also demonstrated a strong relationship with testing futility, $\chi^2(3) = 3962.35$, $p < .001$. The rate of futile testing profiles was similar at follow-up (7.90%). There was a strong association between baseline dementia severity and likelihood of demonstrating a futile testing profile at follow-up, $\chi^2(4) = 1513.40$, $p < .001$. Over 90% of individuals with severe dementia, who were initially able to participate meaningfully testing, no longer could at follow-up. Similarly, there was a strong relationship between baseline MoCA score band and likelihood of demonstrating a futile testing profile at follow-up, $\chi^2(3) = 1627.37$, $p < .001$. Results can help to guide decisions about optimizing use of limited neuropsychological assessment resources.

Among older adults, one of the most common purposes of a neuropsychological evaluation is to ascertain whether there are cognitive impairments indicative of the presence of neurodegenerative disease (Sullivan-Baca, Naylon, Zartman, Ardolf, & Westhafer, 2020). The preponderance of evidence reveals that neuropsychological evaluations are valuable in improving care of older adults by enhancing diagnostic classification and accuracy,

[*]Correspondence should be addressed to Andrew M. Kiselica, 115 Business Loop 70 W, Columbia, MO, USA (akiselica@health.missouri.edu).

as well as assisting in predicting long-term daily life outcomes (Donders, 2020). Neuropsychological evaluation with older adults is a multi-component process, involving collection of data from multiple sources, integrating and analysing that information, drawing clinical conclusions, and communicating impressions and recommendations (Lezak, Howieson, & Loring, 2012). Importantly, neuropsychological testing is only one component of this broader process, and the term testing should not be used synonymously with the term evaluation (American Educational Research Association et al., 2014). Rather, testing refers specifically to the administration of measures to quantify behaviour and psychological processes. Of course, neuropsychologists specialize in the administration, scoring, interpretation, and integration of test data, making testing an integral part of most neuropsychological evaluations (Barth et al., 2003). The usefulness of a neuropsychological testing battery, however, relies in part on a patient's ability to meaningfully participate in testing (Lezak et al., 2012).

Identifying individuals who may not be testable or may yield limited information from testing is important for several reasons. First, putting a patient with cognitive impairments through a potentially frustrating and stressful evaluation when no valuable information will result raises concerns about violating principles of beneficence/nonmaleficence (APA, 2019b). Second, per current billing guidelines (APA, 2019a) one should not charge for testing services when "the patient is … unable to participate in a meaningful way in the testing process; or … there is no expectation that the testing would impact the patient's medical, functional, or behavioural management" (p. 7–8). Third, testability as a clinical construct might have diagnostic and prognostic value in its own right (Benge, Artz, & Kiselica, 2020; Kiselica & Benge, 2019; Pastorek, Hannay, & Contant, 2004). Fourth, testing a person who cannot meaningfully participate places a strain on limited neuropsychological assessment resources, which might be better allocated. Finally, from a research standpoint, there is a need to exclude participants who will not be able to engage meaningfully in cognitive testing if these measures represent outcomes of interest.

Current practice recommendations indicate that a neuropsychologist should proceed with a comprehensive test battery when cognitive screening is positive for possible impairment and there are relevant medical questions that can be answered via testing (Roebuck-Spencer et al., 2017). However, a comprehensive test battery may not always be indicated when significant cognitive impairments limit the utility of information that might be gathered from testing. In such cases, the evaluation may shift from away from an investigative (i.e. testing oriented) approach towards a more supportive (i.e. treatment oriented) one. Certainly, research supports the value of integrating cognitive screening with information on personality, functional, and behavioural status in severely impaired populations (Creavin et al., 2016). Even in the absence of testing, neuropsychological evaluations provide immense value, serving to integrate multiple sources of data to assist with differential diagnosis, prognosis, and treatment planning (Block, Johnson-Greene, Pliskin, & Boake, 2017). Thus, in cases where cognitive impairment is suspected, a neuropsychological evaluation is always indicated, even when comprehensive testing may not be needed.

The issue of inability to participate in neuropsychological testing has not garnered much research attention. One study by Pastorek et al. (2004) reported that 13–19% of patients

in an inpatient traumatic brain injury sample could not meaningfully participate in testing. Notably, inability to participate was actually predictive of long-term outcomes. Additionally, there have been longstanding debates about whether individuals with severe dementia should undergo neuropsychological testing (Boller, Verny, Hugonot-Diener, & Saxton, 2002), but data to support such decisions are limited. Recent research suggests that a significant minority (12–28%) of older adults are unable to tolerate elements of testing (Benge et al., 2020; Kiselica & Benge, 2019; Wolf, Weeda, Wetzels, de Jonghe, & Koopmans, 2019; Wong et al., 2016). However, no research has explored the prevalence of test futility across an entire neuropsychological battery.

Predictors of ability to participate meaningfully in neuropsychological testing

There are typically two main sources of information available to the neuropsychologist regarding whether to proceed with testing. The first source of information is the clinical interview and neurobehavioural status examination. Research indicates that formalized measures of these processes, such as the semi-structured Clinical Dementia Rating (CDR) Dementia Staging Instrument ®, can be used to reliably estimate stages of dementia severity (Morris, 1993, 1997; Morris et al., 1997). Our prior work indicated that this severity status is strongly related to ability to complete measures of executive functioning (Kiselica & Benge, 2019). However, at what point of global dementia severity a neuropsychologist might decide to forego comprehensive testing has not been established.

The second source of information may come from cognitive screening instruments, administered either by a referring provider or the neuropsychologist themselves as part of a neurobehavioural status examination. To our knowledge, no prior studies have examined the relationship between cognitive screening scores and ability to participate meaningfully in subsequent detailed neuropsychological testing. Published studies indicate a moderate-strong correspondence between screening scores and overall dementia severity (Pan et al., 2020; Stewart, Swartz, Tapscott, & Davis, 2019), such that these instruments may be useful for predicting who will demonstrate a futile testing profile. However, this question has yet to be systematically addressed.

A third source of information may come from a prior evaluation. In clinical and research settings, neuropsychologists may be tasked with deciding whether to retest a given individual on the basis of a prior evaluation or they may need to decide whether to recommend retesting at a future time (Lezak et al., 2012). To our knowledge, there is no empirical literature about factors that may signal that a participant/patient will be unable to participate meaningfully in testing in the future, despite being able complete (at least some part) of testing today.

Current study

In light of these gaps in the literature regarding older adults' ability to participate meaningfully in testing, the goals of the current study were as follows:

1. To provide an operational definition for a futile testing profile and examine the base rate of such profiles in an older adult sample that includes individuals with a range of cognitive impairments;

2. To assess the relationship between dementia severity level and the prevalence of futile testing profiles at baseline;

3. To examine the association between cognitive screening scores and probability of demonstrating a futile testing profile at baseline.

4. To examine the relationship of dementia severity and cognitive screening scores with probability of demonstrating a futile testing profile at a follow-up visit.

As we consider these aims, it is important to note that the goal of the research is to assist neuropsychologists and referral sources with clinical decision-making by providing an empirical basis for the choice to forego testing. The results of the current paper should not be construed as a means of limiting a given patient's access to neuropsychological evaluation. Rather, our intent is to assist with clinical judgements regarding the likely futility of proceeding with the comprehensive testing portion of the assessment.

Methods

Sample

We used data from the National Alzheimer's Coordinating Center Uniform Data Set (NACC UDS), which includes information gathered from participants at Alzheimer's Disease Research Centers across the country. Data is collected under the auspices of the IRBs of participating sites and provided to researchers in a deidentified format to protect patient confidentiality. This research was determined to meet exemption from further IRB review under the Human Subjects Determination guidelines at our university. Data were requested through the NACC online portal and provided on October 2, 2020. The dataset included all observations from the onset of data collection through the September 2020 data freeze.

This database included 43,343 individuals with baseline assessments available. We restricted the sample to participants receiving version 3.0 (the most recent version) of the UDS neuropsychological battery (Besser et al., 2018; Weintraub et al., 2018) at baseline ($n = 10,567$). Next, because we were interested in drawing conclusions about the assessment of older adults, we limited the dataset to individuals ages 50+ ($n = 10,238$). Last, since many of the measures in the UDS 3.0 have a significant language component, we restricted the sample to individuals whose primary language was English ($n = 9,263$). The final database included information collected at 40 Research Centers from March of 2015 through September of 2020. Sample characteristics are provided in Table 1.

Measures

Clinical dementia rating (CDR)—The CDR is a structured clinical interview used for severity staging in older adult samples (Morris, 1993). It yields a global rating of cognitive/ functional impairment (0 = cognitively normal, 0.5 = MCI, 1 = mild dementia, 2 = moderate dementia, 3 = severe dementia). For the current study, this measure was used for staging

dementia severity, as it is independent of results from the neuropsychological battery, thus avoiding the possibility of criterion contamination. The CDR has well established psychometric properties (Fillenbaum, Peterson, & Morris, 1996; Morris, 1997; Morris et al., 1997).

Montreal cognitive assessment (MoCA)—The MoCA is a 30-point cognitive screening tool with items assessing language, memory, attention, visuospatial skills, and executive functions (Nasreddine et al., 2005). It is one of the most widely used cognitive screening tools with excellent psychometric properties (Freitas, Prieto, Simões, & Santana, 2014; O'Driscoll & Shaikh, 2017). The MoCA is scored continuously, though cut points are often used to distinguish among individuals with dementia, MCI, and intact cognitive abilities (Milani, Marsiske, Cottler, Chen, & Striley, 2018; Nasreddine et al., 2005). Furthermore, the MoCA website (https://www.mocatest.org/faq/) lists score ranges for different severities, including normal (26–30), mild cognitive impairment (18–25), moderate cognitive impairment (10–17), and severe cognitive impairment (0–9). In our sample, 35.90% of individuals fell in the normal range, 42.80% in the mild range, 12.20% in the moderate range, and 4.80% in the severe range.

UDS 3.0 neuropsychological battery (UDS3NB)—Cognitive tests utilized in the UDS3NB are described in detail in other publications (Besser et al., 2018; Weintraub et al., 2018). Our analyses utilized 12 core scores from this battery (Kiselica, Kaser, Webber, Small, & Benge, 2020; Kiselica, Kaser, et al., 2020; Kiselica Webber, & Benge, 2020a, 2020b). They included 1) three indices of language abilities – The Multilingual Naming Test (Gollan, Weissberger, Runnqvist, Montoya, & Cera, 2012; Ivanova, Salmon, & Gollan, 2013) and the two semantic fluency trials (animals and vegetables); 2) visual test scores – Benson Figure copy and recall trials (Possin, Laluz, Alcantar, Miller, & Kramer, 2011); 3) scores on measures of attention – Number Span Forwards and backwards; 4) indices assessing executive functions – Trail Making Parts A and B (Partington & Leiter, 1949) and a letter fluency test (F- and L-words); and 5) memory test scores – Craft Story immediate and delayed verbatim recall scores (Craft et al., 1996).

Futile testing profile—There is no established definition of what would constitute a futile testing profile. Thus, we operationalized this concept in a number of ways. First, we considered testing to be futile if the overwhelming majority of tests were unable to be completed due to cognitive or behavioural factors. Second, a profile that yielded extremely low scores, without any meaningful variation, would be unlikely to provide incrementally useful information beyond that gathered in an interview (Lezak et al., 2012). Thus, we defined futile testing profiles as those that included   75% of tests that met at least one of the following criteria:

1. Unable to be completed due to a cognitive/behavioural problem.

2. A score at the floor of the range of possible scores. For most tests, the lowest possible score was 0, with the exception of Trail Making Parts A and B. Maximum score was 150 seconds for Part A and 300 seconds for Part B.

3.   Performance below the second percentile based on demographically adjusted norms for the UDS3NB (Weintraub et al., 2018), corresponding to an extremely low score (Guilmette et al., 2020).

Analyses

Baseline analyses—First, we present descriptive information of the percentage of individuals meeting the different futile test performance criteria by each individual test in the UDS3NB. Second, we examined the proportion of individuals in the overall sample who demonstrated a futile test profile. Third, we assessed rates of futile test profiles by CDR severity status and MoCA ranges, using chi-square tests of independence to compare proportions across groups. Finally, we investigated diagnostic accuracy for different MoCA score cut-offs used to predict futile test profiles. Specifically, we constructed a receiver operating characteristic (ROC) curve and examined sensitivity, specificity, positive predictive value, negative predictive value, and post-test probabilities, following recommendations of Smith, Ivnik, and Lucas (2008). Given the wealth of research showing the incremental utility of neuropsychological assessments in improving patient care (Donders, 2020), the baseline assumption should be that comprehensive testing would benefit all patients. Thus, there should a high threshold for foregoing testing and a corresponding preference for high specificity and positive predictive value for procedures used to determine who is likely to demonstrate a futile testing profile.

Follow-up analyses—We used the first follow-up data (time 2; occurred ~1 year after baseline) to examine likelihood of being able to meaningfully participate in retesting. Specifically, we assess the relationship of baseline cognitive severity status and cognitive screening scores with likelihood of developing a futile testing profile at follow-up (repeating baseline analyses, using follow-up data as outcomes). Analyses excluded individuals with futile test profiles at baseline.

Results

Futile test profiles at baseline

The percentage of individuals meeting the different futile test performance criteria by each individual test in the UDS3NB are provided in Table 2. At least one of these criteria was met for ≥ 75% of measures in 7.40% of the sample ($n = 686$); these individuals were considered to have produced futile testing profiles.

There was a strong relationship between dementia severity as indexed by the CDR and likelihood of demonstrating a futile testing profile, $\chi^2(4) = 3559.77$, $p < .001$ (see Figure 1). Futile testing profiles were exceedingly rare in individuals rated as cognitively normal (0.0002%) but increased with increasing dementia severity. However, it was only in the severe range (CDR = 3.0) of dementia that the majority of participants (91.10%) produced futile profiles.

Similarly, there was a strong relationship between MoCA score range and likelihood of demonstrating a futile testing profile, $\chi^2(3) = 3962.35$, $p < .001$ (see Figure 1). No individuals scoring 26–30 on the MoCA demonstrated a futile testing profile; however,

likelihood of a futile test profile increased as MoCA score fell, with 71.10% of individuals scoring 0–9 points meeting futility criteria.

A ROC curve with MoCA scores predicting futile test profiles yielded an excellent *AUC* value of .97, [CI$_{95\%}$ = [.97, .98]. Diagnostic statistics for cut points used to identify individuals likely to have futile test profiles are summarized in Table 3. Scores on the MoCA below 7 yielded perfect specificity and a post-test probability of a futile test profile of .86.

Futile test profiles at time 2

At first follow-up (time 2), there were 4,589 individuals with sufficient data available to assess for futile test profiles. In this group, there 364 (7.9%) individuals demonstrated futile testing profiles. There was a strong association between time 1 dementia severity and likelihood of demonstrating a futile testing profile at time 2, $\chi^2(4) = 1513.40$, $p <$ .001 (see Figure 2). Virtually no individuals coded as cognitively normal at time 1 (0.02%) demonstrated a futile testing profile at time 2, but this proportion increased with severity status, reaching 92.90% for individuals with severe dementia at time 1.

Similarly, there was a strong relationship between baseline MoCA score at time 1and likelihood of demonstrating a futile testing profile at time 2, $\chi^2(3) = 1627.37$, $p < .001$ (see Figure 2). Virtually no one with a baseline MoCA score of 26–30 at time 1 (0.10%) demonstrated a futile testing profile at time 2, but this proportion increased as MoCA score decreased, reaching 77.00% for individuals scoring 0–9 points at time 1.

A ROC curve with baseline MoCA scores predicting follow-up futile test profiles yielded an excellent *AUC* value of .95, [CI$_{95\%}$ = [.94, .96]. Diagnostic statistics for cut points used to identify individuals likely to have futile test profiles are summarized in Table 3. Scores on the MoCA below 9 yielded perfect specificity and a post-test probability of a futile test profile of .84.

Discussion

The goals of this manuscript were to assess the prevalence of futile testing profiles and examine the associations of dementia severity and cognitive screening scores with testing futility at baseline and follow-up evaluations. Results have implications for practitioners and researchers in the neurosciences who are involved in decision-making regarding the appropriateness of patients/participants for comprehensive neuropsychological testing.

Prevalence of futile testing profiles at baseline

We first investigated the prevalence of meeting different futile testing profile criteria for each of the core scores in the UDS3NB (see Table 2). Several important findings emerged from this examination. First, being unable to complete specific neuropsychological tests due to a cognitive/behavioural problem was relatively rare, except in the case of Trail Making Part B (occurred in 10.70% of the sample). Results replicate prior findings, which suggested Trail Making Part B is one of the more difficult to complete tests for cognitively compromised individuals (Kiselica & Benge, 2019; Wong et al., 2016). In fact, the inability to complete

Trail Making Part B was a strong predictor in its own right of increasing dependence in day to day functioning in recent analyses (Benge et al., 2020).

Relatedly, our findings suggested that performance at the floor level of neuropsychological tests was also relatively rare, with the exception of recall scores (10.10–10.40%). This result makes sense, as most neuropsychological tests in this geriatric focused battery require a very basic response to obtain minimal points (e.g. saying one word on letter fluency; drawing one design element on the Benson Figure); that is, if an individual is able to participate in the testing at all, he/she will be unlikely to score at the floor. However, recall measures require intact consolidation abilities, such that those who are densely amnestic will not be able to give a single scorable answer even when they can participate (Heilman & Valenstein, 2010). That being said, producing extremely low normed scores on individuals tests is relatively common (4.30–22.80%, depending on the task), conforming with a wealth of prior research on the base rates of low scores (Binder, Iverson, & Brooks, 2009; Brooks & Iverson, 2010; Brooks, Iverson, & Holdnack, 2013; Brooks, Iverson, Holdnack, & Feldman, 2008; Brooks, Iverson, & White, 2007; Holdnack et al., 2017; Kiselica, Kaser, et al., 2020; Kiselica Webber et al., 2020a).

In contrast, being unable to complete testing or having floor or low scores across the majority of the battery was fairly rare (7.40%). This was the first study to our knowledge that examined the rate of futile testing profiles in older adults, and our results suggest that most older adults can participate meaningfully in cognitive testing. However, it must be noted that the rate of testing futility may change, depending on the makeup of the sample under study. For example, the prevalence of dementia in the current sample was 16.76%, much higher than in the general U.S. population (8.50%; Prince et al., 2013), but much lower than that reported at some memory clinics (36%; Fischer et al., 2009). Thus, the rate of futile test profiles found in our sample likely overestimates that which would be found in samples derived from the general population but underestimates that which would be found in some clinic settings. Further research is needed to assess differences in rates of test futility across assessment contexts.

Another point to consider here is that the operational definition of a futile testing profile is ultimately arbitrary. Of course, this definition may vary by provider preferences and idiosyncrasies of the particular neuropsychological test battery that is administered. However, we felt that having 75% of test data not be useful was a conservative definition of test futility that could be applied across most assessment contexts. Regardless, this research provides a useful starting point for further investigations into the inability to meaningfully engage in testing.

Futile testing profiles by dementia severity status at baseline

Results indicated a strong relationship between severity status (as judged by the Clinical Dementia Rating global score) and prevalence of futile testing profiles. Indeed, in patients without dementia, testing was almost never futile (0.0002% among cognitively normal individuals, and 3.10% among individuals with MCI). However, as dementia severity increased, test futility became more and more likely, being the rule rather than the exception among individuals with severe dementia (91.10%). Findings imply that in cases

of severe dementia identified through clinical interview, collateral report, record review, etc., administering a typical neuropsychological battery is unlikely to yield useful information. It should be noted that there are alternative screening measures for severely impaired individuals, such as the Severe Impairment Battery and the Dementia Rating Scale (Jurica, Leitten, & Mattis, 1988; Panisset, Roudier, Saxton, & Boiler, 1994), and using such instruments as indicated by clinical interview/judgement may be warranted.

Of course, the inability to complete neuropsychological tests may provide useful information in its own right (Benge et al., 2020; Pastorek et al., 2004), and a minority of individuals with severe dementia will be able to participate meaningfully in testing. Furthermore, our clinical experience suggests that sometimes trying and failing to produce valid responses on simple tasks "hits home" the need for intervention, support, or placement to family members and referring providers, who may not have initially appreciated the severity of cognitive decline. Thus, there continues to be a need for judgement by individual providers about whether a standard neuropsychological test battery should be administered.

Again, based on current practice recommendations (APA, 2019a), as well as our clinical experience, this judgement call hinges on one important question: "Are the test results likely to meaningfully change the care of this patient?" For instance, imagine a patient with an established dementia diagnosis and a family who understands and accepts her condition. She has been placed in an excellent skilled nursing facility and is followed by a well-regarded geriatrician. In such a case, testing is unlikely to change the plan of care much beyond what might be suggested following a careful neurobehavioural status examination and brief cognitive screening. On the other hand, for cases where the severity of dementia is unclear, there is a need to justify a higher level of care, or there are important medicolegal questions (e.g. does this patient have capacity to consent to a needed surgery?) at play, a comprehensive test battery is likely to yield greater benefit.

Futile testing profiles by MoCA ranges at baseline

Beyond the clinical interview, a clinician or researcher might be interested in using a cognitive screening instrument, such as the MoCA, to assist with the decision of whether to test. Our results suggest that the MoCA is not as informative as the CDR for determining who will produce a futile profile. However, results provide evidence of the MoCA as adjunctive evidence for decisions to forego testing. Near floor performance (<7 points) on the MoCA indicates a very high post-test probability (.86) that an individual will demonstrate a futile testing profile (see Table 3). Such a low score only occurred in 1.70% of our sample. Notably, 5 points on the MoCA is the recommended cut point for a severe dementia diagnosis (Pan et al., 2020), suggesting converging evidence that indicators of severe dementia may meaningfully predict who will produce a futile profile.

It must be noted that the post-test probabilities for the MoCA cut points in predicting testing futility are influenced by the base rate (i.e. the pre-test probability) of this phenomenon. In our sample, the prevalence of futile test profiles was fairly low (7.40%). In dementia clinics, the base rate of futile testing could be much higher, given that dementia prevalence can be more than two times as high as in our sample in such settings (Fischer et al., 2009). If the base rate of futile test profiles also doubled, the post-test probability of a futile test given a

MoCA score < 7 would jump to .93. In contrast, if the base rate was halved, as one might expect in the general population (Prince et al., 2013), the post-test probability of a futile test given a MoCA score < 5 would drop to .75. Thus, the utility of the MoCA for predicting who is appropriate for testing will depend greatly on the population in which this screener is employed.

Another important point to consider is that there may be other relevant sources of information available to the neuropsychologist when considering the appropriateness of testing for a given patient/participant. For example, neuroimaging may find diffuse brain injuries, which could suggest that testing could be difficult. Similarly, the medical record may have notes from other providers who have struggled to obtain meaningful information from the individual.

Predicting futile test profiles at time 2

Among individuals who could participate meaningfully in testing at baseline, 7.90% had a futile testing profile at time 2 (occurred about a year later). Similar to the baseline findings, baseline cognitive severity status was strongly related to test futility at follow-up, with over 90% of those with severe dementia no longer able to participate meaningfully in testing at time 2. Time 1 MoCA screening scores were also strongly associated with test futility at follow-up, with individuals scoring below 9 have a strong post-test probability (.84) of a futile test profile at follow-up. In summary, findings closely mirrored those for the baseline data. This is the first study, to our knowledge, to provide empirical data on predicting the ability to retest among older adults. Of course, the decision to retest is made based on a number of factors beyond severity of impairment, including suspected diagnosis, anticipated prognosis, and availability of treatment resources. Clearly, more research is needed to account for these many factors and help neuropsychologists make the decision about when to re-evaluate.

Notably, the longitudinal predictive utility of cognitive severity status and cognitive screening scores may have been attenuated somewhat by selective attrition. Post hoc analyses revealed that individuals with sufficient follow-up data ($M = 1.48$, $SD = 2.40$) were significantly less impaired than those without sufficient data ($M = 2.50$, $SD = 3.82$) on the Clinical Dementia Rating sum of box scores at baseline, $t(7884.73) = 15.36$, $p < .001$, $d = .32$. A similar result was found for the MoCA, $t(8460.25) = -12.59$, $p < .001$, $d = .27$, with higher baseline scores for those with sufficient follow-up data ($M = 23.22$, $SD = 5.38$) versus those without ($M = 21.62$, $SD = 6.49$). Thus, if all individuals could have been included at time 2, the proportion of individuals with futile profiles may have been higher and prediction results may have been stronger.

Limitations

In addition to the possibility of differential attrition, there are other important factors that impact interpretation of the current results. First, the UDS3NB was designed specifically for clinical research with older adults and is relatively brief (approximately 20–40 minutes in our experience). While there are a growing number of tools to support its application in clinical samples (Devora, Beevers, Kiselica, & Benge, 2020; Kiselica, Kaser, et al.,

2020; Kiselica Webber et al., 2020a; Kiselica, Kaser, et al., 2020; Kiselica Webber, & Benge, 2020b; Kiselica, Kaser, et al., 2020; Sachs et al., 2020; Weintraub et al., 2018), it is not as lengthy or extensive as other more commonly utilized batteries that can last several hours (Rabin, Paolillo, & Barr, 2016). Thus, our findings may not generalize to other batteries commonly employed in clinical practice, and they highlight the importance of tailoring testing strategies to the population of interest. Second, we focused on the MoCA as cognitive screener because it is given in the current version of the UDS and appears to be increasingly preferred over other measures in assessments of older adults (Ciesielska et al., 2016; Saczynski et al., 2015; Weintraub et al., 2018). However, it must be acknowledged that other tools, like the Mini-Mental Status Exam (Folstein, Folstein, & McHugh, 1975), could be used for the purpose of assessing who is appropriate for testing. Future research could compare several potential screeners for this purpose. Finally, the UDS sample tends to skew white and highly educated, limiting generalizability of findings in more diverse groups. However, recent research is beginning to address this lack of representation (Sachs et al., 2020), and our results may soon be replicated in more diverse samples.

Conclusions

To our knowledge, this study is the first to investigate the prevalence of futile testing profiles among older adults. We found that the vast majority of participants in our sample were able to meaningfully participate in testing. Individuals with severe dementia were at the highest risk of demonstrating futile testing profiles at baseline and follow-up (>90%). Notably, cognitive screening scores do not appear to yield much additional information on who is likely to demonstrate a futile profile beyond classifying severity status, though a low MoCA score (<7 for predicting baseline futility and < 9 for predicting follow-up futility) suggests a high likelihood of futile testing. The research thus provides some heuristics to assist practitioners in deciding who is suitable to test; however, it is ultimately the decision of the individual researcher or provider as to whether testing is indicated. The default position should always be to test, given the clearly established utility of neuropsychological evaluations (Donders, 2020), with testing deferred only in the rare cases where these procedures have a high cost-benefit ratio. In closing, we feel it important to reiterate that these results should not be construed to deny anyone neuropsychological evaluation services and are intended for the purposes of improving clinical decision-making regarding the utility of proceeding with testing or retesting.

Acknowledgements

Craft, PhD), P50 AG005136 (PI Thomas Grabowski, MD), P30 AG062715-01 (PI Sanjay Asthana, MD, FRCP), P50 AG005681 (PI John Morris, MD), P50 AG047270 (PI Stephen Strittmatter, MD, PhD).

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014) The standards for educational and psychological testing. Washington, D.C.: AERA.

APA. (2019a) Ethical principles of psychologists and code of conduct. Washington, D.C.: American Psychological Association.

APA. (2019b) Psychological and Neuropsychological Testing Billing and Coding Guide. Washington, D.C.: American Psychological Association.

Barth JT, Pliskin N, Axelrod B, Faust D, Fisher J, Harley JP, … Silver C (2003). Introduction to the NAN 2001 Definition of a Clinical Neuropsychologist NAN Policy and Planning Committee. Archives of Clinical Neuropsychology, 18(5), 551–555. [PubMed: 14591449]

Benge JF, Artz JD, & Kiselica AM (2020). The ecological validity of the Uniform Data Set 3.0 neuropsychological battery in individuals with mild cognitive impairment and dementia. The Clinical Neuropsychologist, 1–18. 10.1080/13854046.2020.1837246. [Epub Ahead of Print].

Besser L, Kukull W, Knopman DS, Chui H, Galasko D, Weintraub S, … Morris JC (2018). Version 3 of the national Alzheimer's coordinating center's Uniform Data Set. Alzheimer Disease and Associated Disorders, 32, 351. [PubMed: 30376508]

Binder LM, Iverson GL, & Brooks BL (2009). To err is human:"Abnormal" neuropsychological scores and variability are common in healthy adults. Archives of Clinical Neuropsychology, 24 (1), 31–46. [PubMed: 19395355]

Block CK, Johnson-Greene D, Pliskin N, & Boake C (2017). Discriminating cognitive screening and cognitive testing from neuropsychological assessment: implications for professional practice. The Clinical Neuropsychologist, 31, 487–500. [PubMed: 27937143]

Boller F, Verny M, Hugonot-Diener L, & Saxton J (2002). Clinical features and assessment of severe dementia. A Review. European Journal of Neurology, 9, 125–136. 10.1046/j.1468-1331.2002.00356.x [PubMed: 11882053]

Brooks BL, & Iverson GL (2010). Comparing actual to estimated base rates of "Abnormal" scores on neuropsychological test batteries: Implications for interpretation [Article]. Archives of Clinical Neuropsychology, 25(1), 14–21. 10.1093/arclin/acp100 [PubMed: 20015965]

Brooks BL, Iverson GL, & Holdnack JA (2013). Understanding and using multivariate base rates with the WAIS-IV/WMS-IV. Cambridge, MA: Elsevier Academic Press Inc. 10.1016/b978-0-12-386934-0.00002-x

Brooks BL, Iverson GL, Holdnack J, & Feldman HH (2008). Potential for misclassification of mild cognitive impairment: A study of memory scores on the Wechsler Memory Scale-III in healthy older adults [Article]. Journal of the International Neuropsychological Society, 14(3), 463–478. 10.1017/s1355617708080521 [PubMed: 18419845]

Brooks BL, Iverson GL, & White T (2007). Substantial risk of "Accidental MCI" in healthy older adults: Base rates of low memory scores in neuropsychological assessment [Article]. Journal of the International Neuropsychological Society, 13, 490–500. 10.1017/d1355617707070531 [PubMed: 17445298]

Ciesielska N, Sokolowski R, Mazur E, Podhorecka M, Polak-Szabela A, & Kedziora-Kornatowska K (2016). Is the montreal cognitive assessment (MoCA) test better suited than the Mini-Mental State Examination (MMSE) in mild cognitive impairment (MCI) detection among people aged over 60? Meta-analysis. Psychiatria Polska, 50, 1039–1052. [PubMed: 27992895]

Craft S, Newcomer J, Kanne S, Dagogo-Jack S, Cryer P, Sheline Y, … Alderson A (1996). Memory improvement following induced hyperinsulinemia in Alzheimer's disease. Neurobiology of Aging, 17(1), 123–130. [PubMed: 8786794]

Creavin ST, Wisniewski S, Noel-Storr AH, Trevelyan CM, Hampton T, Rayment D, …Cullum S. (2016). Mini-Mental State Examination (MMSE) for the detection of dementia in clinically unevaluated people aged 65 and over in community and primary care populations. Cochrane Database of Systematic Reviews, (1), Article Cd011145.10.1002/14651858.CD011145.pub2

Devora PV, Beevers S, Kiselica AM, & Benge JF (2020). Normative data for derived measures and discrepancy scores for the Uniform Data Set 3.0 Neuropsychological Battery. Archives of Clinical Neuropsychology, 35(1), 75–89.

Donders J (2020). The incremental value of neuropsychological assessment: A critical review. The Clinical Neuropsychologist, 34(1), 56–87. [PubMed: 31014166]

Fillenbaum G, Peterson B, & Morris J (1996). Estimating the validity of the Clinical Dementia Rating scale: the CERAD experience. Aging Clinical and Experimental Research, 8, 379–385.

Fischer C, Yeung E, Hansen T, Gibbons S, Fornazzari L, Ringer L, & Schweizer T (2009). Impact of socioeconomic status on the prevalence of dementia in an inner city memory disorders clinic. International Psychogeriatrics, 21, 1096. [PubMed: 19712540]

Folstein MF, Folstein SE, & McHugh PR (1975). "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research, 12, 189–198. [PubMed: 1202204]

Freitas S, Prieto G, Simões MR, & Santana I (2014). Psychometric properties of the Montreal Cognitive Assessment (MoCA): an analysis using the Rasch model. The Clinical Neuropsychologist, 28(1), 65–83. [PubMed: 24404822]

Gollan TH, Weissberger GH, Runnqvist E, Montoya RI, & Cera CM (2012). Self-ratings of spoken language dominance: A Multilingual Naming Test (MINT) and preliminary norms for young and aging Spanish-English bilinguals. Bilingualism: Language and Cognition, 15, 594–615.

Guilmette TJ, Sweet JJ, Hebben N, Koltai D, Mahone EM, Spiegler BJ, … Participants C (2020). American academy of clinical neuropsychology consensus conference statement on uniform labeling of performance test scores. The Clinical Neuropsychologist, 34, 437–453. [PubMed: 32037942]

Heilman MKM, & Valenstein E (2010). Clinical neuropsychology. Oxford: Oxford University Press.

Holdnack JA, Tulsky DS, Brooks BL, Slotkin J, Gershon R, Heinemann AW, & Iverson GL (2017). Interpreting patterns of low scores on the NIH toolbox cognition battery [Article]. Archives of Clinical Neuropsychology, 32, 574–584. 10.1093/arclin/acx032 [PubMed: 28419177]

Ivanova I, Salmon DP, & Gollan TH (2013). The multilingual naming test in Alzheimer's disease: clues to the origin of naming impairments. Journal of the International Neuropsychological Society, 19, 272–283. [PubMed: 23298442]

Jurica PJ, Leitten CL, & Mattis S (1988). DRS-2 dementia rating scale-2: professional manual. Lutz, FL: Psychological Assessment Resources.

Kiselica AM, & Benge JF (2019). Quantitative and qualitative features of executive dysfunction in Frontotemporal and Alzheimer's Dementia. Applied Neuropsychology: Adult. 10.1080/23279095.2019.1652175. Epub ahead of print.

Kiselica AM, Kaser A, Webber T, Small B, & Benge J (2020). Development and preliminary validation of standardized regression-based change scores as measures of transitional cognitive decline. Archives of Clinical Neurospychology, 35(7), 1168–1181. 10.1093/arclin/acaa042

Kiselica AM, Webber T, & Benge J (2020a). Using multivariate base rates of low scores to understand early cognitive declines on the uniform data set 3.0 Neuropsychological Battery. Neuropsychology, 34, 629–640. [PubMed: 32338945]

Kiselica AM, Webber TA, & Benge JF (2020b). The uniform data set 3.0 Neuropsychological Battery: Factor structure, invariance testing, and demographically-adjusted factor score calculation. Journal of the International Neuropsychological Society, 26, 576–586. 10.1017/S135561772000003X [PubMed: 32063246]

Lezak M, Howieson D, & Loring D (2012). Neuropsychological assessment, 5th ed. New York, NY: Oxford University Press.

Milani SA, Marsiske M, Cottler LB, Chen X, & Striley CW (2018). Optimal cutoffs for the Montreal Cognitive Assessment vary by race and ethnicity. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 10, 773–781.

Morris JC (1993). The clinical dementia rating (CDR): Current version and scoring rules. Neurology, 43, 2412.

Morris JC (1997). Clinical dementia rating: a reliable and valid diagnostic and staging measure for dementia of the Alzheimer type. International Psychogeriatrics, 9(S1), 173–176. [PubMed: 9447441]

Morris JC, Ernesto C, Schafer K, Coats M, Leon S, Sano M, … Woodbury P (1997). Clinical dementia rating training and reliability in multicenter studies: the Alzheimer's Disease Cooperative Study experience. Neurology, 48, 1508–1510. [PubMed: 9191756]

Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, … Chertkow H (2005). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. Journal of the American Geriatrics Society, 53, 695–699. [PubMed: 15817019]

O'Driscoll C, & Shaikh M (2017). Cross-cultural applicability of the Montreal Cognitive Assessment (MoCA): a systematic review. Journal of Alzheimer's Disease, 58, 789–801.

Pan IMY, Lau MS, Mak SC, Hariman KW, Hon SKH, Ching WK, … Chan C (2020). Staging of Dementia Severity With the Hong Kong Version of the Montreal Cognitive Assessment (HK-MoCA)'s. Alzheimer Disease and Associated Disorders, 34, 333–338. [PubMed: 32701514]

Panisset M, Roudier M, Saxton J, & Boiler F (1994). Severe impairment battery: a neuropsychological test for severely demented patients. Archives of Neurology, 51(1), 41–45. [PubMed: 8274108]

Partington JE, & Leiter RG (1949). Partington pathways test. Psychological Service Center Journal, 1, 11–20.

Pastorek NJ, Hannay HJ, & Contant CS (2004). Prediction of global outcome with acute neuropsychological testing following closed-head injury. Journal of the International Neuropsychological Society: JINS, 10, 807. [PubMed: 15637771]

Possin KL, Laluz VR, Alcantar OZ, Miller BL, & Kramer JH (2011). Distinct neuroanatomical substrates and cognitive mechanisms of figure copy performance in Alzheimer's disease and behavioral variant frontotemporal dementia. Neuropsychologia, 49 (1), 43–48. [PubMed: 21029744]

Prince M, Bryce R, Albanese E, Wimo A, Ribeiro W, & Ferri CP (2013). The global prevalence of dementia: a systematic review and metaanalysis. Alzheimer's & Dementia, 9(1), 63–75. e62.

Rabin LA, Paolillo E, & Barr WB (2016). Stability in test-usage practices of clinical neuropsychologists in the United States and Canada over a 10-year period: A follow-up survey of INS and NAN members. Archives of Clinical Neuropsychology, 31, 206–230. [PubMed: 26984127]

Roebuck-Spencer TM, Glen T, Puente AE, Denney RL, Ruff RM, Hostetter G, & Bianchini KJ (2017). Cognitive screening tests versus comprehensive neuropsychological test batteries: A national academy of neuropsychology education paper. Archives of Clinical Neuropsychology, 32, 491–498. 10.1093/arclin/acx021 [PubMed: 28334244]

Sachs BC, Steenland K, Zhao L, Hughes TM, Weintraub S, Dodge HH, … Goldstein FC (2020). Expanded demographic norms for version 3 of the Alzheimer disease centers' neuropsychological test battery in the uniform data set. Alzheimer Disease and Associated Disorders, 34, 191–197. 10.1097/wad.0000000000000388 [PubMed: 32483017]

Saczynski JS, Inouye SK, Guess J, Jones RN, Fong TG, Nemeth E, … Marcantonio ER (2015). The Montreal cognitive assessment: Creating a crosswalk with the mini-mental state examination [Article]. Journal of the American Geriatrics Society, 63, 2370–2374. 10.1111/jgs.13710 [PubMed: 26503296]

Smith GE, Ivnik RJ, & Lucas J (2008). Assessment techniques: Tests, test batteries, norms, and methodological approaches. Textbook of Clinical Neuropsychology, 38–57. New York, NY: Taylor & Francis.

Stewart P, Swartz J, Tapscott B, & Davis B (2019). C-36 montreal cognitive assessment for dementia severity rating in a diagnostically heterogeneous clinical cohort. Archives of Clinical Neuropsychology, 34, 1065.

Sullivan-Baca E, Naylon K, Zartman A, Ardolf B, & Westhafer JG (2020). Gender differences in veterans referred for neuropsychological evaluation in an outpatient neuropsychology consultation service. Archives of Clinical Neuropsychology, 35(5), 562–575. 10.1093/arclin/acaa008 [PubMed: 32128586]

Weintraub S, Besser L, Dodge HH, Teylan M, Ferris S, Goldstein FC, … Morris JC (2018). Version 3 of the Alzheimer disease centers' neuropsychological test battery in the uniform data set (UDS). Alzheimer Disease and Associated Disorders, 32(1), 10. [PubMed: 29240561]

Wolf ET, Weeda WD, Wetzels RB, de Jonghe JFM, & Koopmans RCTM (2019). Course of cognitive functioning in institutionalized persons with moderate to severe dementia: Evidence from the severe impairment battery short version. Journal of the International Neuropsychological Society, 25, 204–214. 10.1017/s1355617718000991 [PubMed: 30457078]

Wong S, Bertoux M, Savage G, Hodges JR, Piguet O, & Hornberger M (2016). Comparison of prefrontal atrophy and episodic memory performance in dysexecutive Alzheimer's disease and behavioral-variant frontotemporal dementia [Article]. Journal of Alzheimers Disease, 51, 889–903. 10.3233/jad-151016
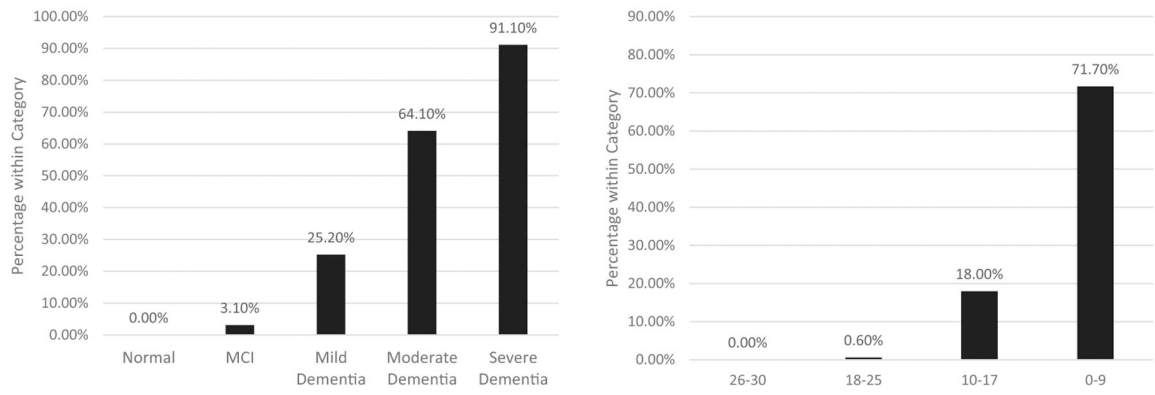
Figure 1.
Percentage of individuals with futile test profiles at baseline by baseline Clinical Dementia Rating severity status (left panel) and Montreal Cognitive Assessment range (right panel).
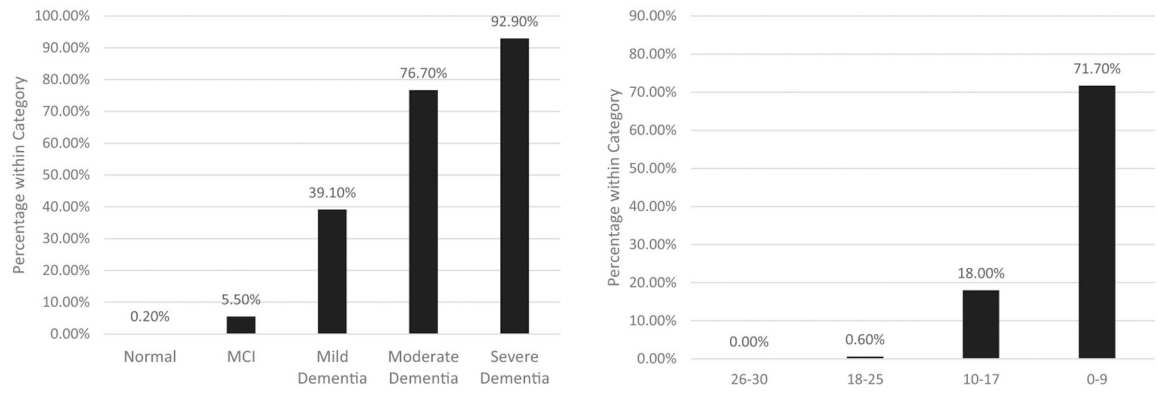
Figure 2.
Percentage of individuals with futile test profiles at time 2 (~1-year follow-up) by time
1 (baseline) Clinical Dementia Rating severity status (left panel) and Montreal Cognitive
Assessment range (right panel).

Table 1.

Sample demographic characteristics by severity stage (based on the clinical dementia rating) at baseline

		Normal	MCI	Mild dementia	Moderate dementia	Severe dementia	Total sample
	M (SD) or %						
N		4,205	3,505	1,092	317	144	9,263
Age		69.69 (8.00)	71.09 (8.41)	70.26 (9.49)	70.37 (10.74)	73.28 (10.31)	70.37 (8.52)
Education		16.36 (2.50)	15.99 (2.80)	15.56 (2.86)	15.40 (2.71)	15.68 (2.88)	16.09 (2.69)
% Female		65.90%	50.70%	48.20%	56.50%	52.80%	57.50%
Race							
White		77.10%	82.60%	88.90%	84.90%	88.20%	81.00%
Black		18.90%	14.10%	7.30%	12.90%	9.00%	15.40%
Asian		1.90%	1.70%	1.80%	1.30%	2.10%	1.80%
Native American/Alaska Native		1.10%	0.80%	0.50%	0.30%	0.00%	0.90%
Hawaiian/Pacific Islander		0.10%	0.10%	0.30%	0.00%	0.00%	0.10%
Other		0.50%	0.30%	0.50%	0.30%	0.00%	0.40%
Ethnicity							
Non-Hispanic		95.70%	96.80%	97.70%	97.80%	97.20%	96.20%
Hispanic		4.20%	2.70%	2.20%	2.20%	2.80%	3.30%

Table 2.

Percentage of the sample meeting testing futility criteria for each test in the uniform data set 3.0 neuropsychological battery at baseline, along with other reasons for missing data

	Test futility criteria			Other reasons for missing data		
	Unable to be completed due to cognitive/behavioural problem	Floor performance	Extremely low score	Unable to complete due to physical problem	Declined to participate	Other problem/missing
Benson figure copy	2.20%	0.60%	11.90%	0.40%	0.30%	1.80%
Benson figure recall	2.80%	10.10%	22.80%	0.30%	0.50%	3.30%
Animal fluency	2.00%	0.30%	14.40%	0.10%	0.30%	1.30%
Vegetable fluency	2.20%	1.30%	14.90%	0.10%	0.30%	1.70%
Trailmaking Part A	3.50%	3.00%	18.40%	0.50%	0.40%	2.10%
Trailmaking Part B	10.70%	6.90%	14.40%	0.60%	1.20%	2.50%
Letter fluency	2.40%	0.20%	9.40%	0.10%	0.60%	5.40%
Craft story immediate	2.40%	1.70%	17.40%	0.20%	0.40%	2.60%
Craft story delay	2.80%	10.40%	21.20%	0.20%	0.50%	2.70%
Number span forward	1.80%	0.30%	4.30%	0.10%	0.20%	2.10%
Number span backward	2.00%	1.70%	8.50%	0.10%	0.30%	2.10%
Multilingual naming test	2.60%	0.20%	15.60%	0.20%	0.40%	2.40%

Table 3.

Diagnostic accuracy statistics for baseline montreal cognitive assessment (MoCA) score cutpoints used to identify individuals likely to have futile test profiles at baseline and time 2

Baseline

MoCA cut point	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Post-test probability
<10	.58	.99	.76	.97	.76
<9	.51	.99	.82	.96	.81
<8	.42	.99	.82	.95	.82
<7	.34	1.00	.86	.95	.86

Time 2

MoCA cut point	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Post-test probability
<10	.33	.99	.78	.95	.79
<9	.26	1.00	.84	.94	.84