Harmonizing neuropsychological assessment for mild neurocognitive disorders in Europe


DOI: 10.1002/alz.12365
Abstract

Introduction: Harmonized neuropsychological assessment for neurocognitive disorders, an international priority for valid and reliable diagnostic procedures, has been achieved only in specific countries or research contexts.

Methods: To harmonize the assessment of mild cognitive impairment in Europe, a workshop (Geneva, May 2018) convened stakeholders, methodologists, academic, and non-academic clinicians and experts from European, US, and Australian harmonization initiatives.

Results: With formal presentations and thematic working-groups we defined a standard battery consistent with the U.S. Uniform DataSet, version 3, and homogeneous methodology to obtain consistent normative data across tests and languages. Adaptations consist of including two tests specific to typical Alzheimer’s disease and behavioral variant frontotemporal dementia. The methodology for harmonized normative data includes consensus definition of cognitively normal controls, classification of confounding factors (age, sex, and education), and calculation of minimum sample sizes.

Discussion: This expert consensus allows harmonizing the diagnosis of neurocognitive disorders across European countries and possibly beyond.

Keywords: Alzheimer’s disease, cognitive assessment, diagnosis, mild cognitive impairment, mild neurocognitive disorders, standard neuropsychological assessment
INTRODUCTION

A key early step in the clinical diagnostic process for persons with cognitive complaints who are referred to memory clinics consists of ascertaining the presence of objective cognitive impairment \(^{1,2}\) by examining performance on a set of neuropsychological tests. As shown in Supplemental Table S1, different tests are more sensitive to different disorders. Heterogeneous batteries could therefore result in different diagnoses for patients. Reliable clinical actions require that different diagnostic procedures operationalize the definition of the target clinical disorder consistently. One way to accomplish this is by consistent test selection. Such standard procedures would enable the use of biomarkers and treatment in line with their demonstrated informative or therapeutic value, and uniformly across centers. In the same way, valid research procedures require consistent operationalization of the same clinical construct across research settings. It is therefore desirable that the selection of patients who are eligible for the full diagnostic procedure or for research studies be based on a standard common neuropsychological assessment, operationalizing the target condition consistently.\(^ {3,4}\) Post hoc computations permit alignment of scores from heterogeneous batteries and pooling of data from different centers for research aims\(^ {5,6}\) thereby facilitating analyses of large multi-site data sets. Such computations, however, cannot amend the upstream inclusion of heterogeneous patients.

Many efforts have been made to tackle this problem, providing resources to support harmonization (Supplemental Table S2). For example, many US research centers have been able to standardize neuropsychological assessments on a large scale.\(^ {7}\) Similarly, German-speaking countries widely adopted the Consortium to Establish a Registry for Alzheimer’s Disease–Neuropsychological Assessment Battery (CERAD-NAB) for the diagnosis of patients with dementia (Supplemental Table S2), and recently, a Chinese effort defined a standard battery for clinical use.\(^ {8}\) This work aims toward standardizing neuropsychological assessments for detecting mild cognitive impairment (MCI) consistently in people attending European memory clinics with cognitive complaints. We leveraged previous initiatives and incorporated the complementary expertise of academic and non-academic clinicians to provide standard procedures that will reduce costs and effort in clinics and benefit research activities.

METHODS

This initiative follows the Strategic Biomarker Roadmap, a methodological framework specific to biomarker validation, adapted from oncology to the field of dementia. This framework outlined the appropriate sequence of validation steps for diagnostic biomarkers, and the priority of standardizing neuropsychological assessment as a prerequisite for their proper validation.\(^ {4,9}\) Because many studies of clinical validity and utility are performed on patients from memory clinics, harmonizing neuropsychological assessment for the clinical settings would have double benefit, improving research as well as clinical procedures at once.

A workshop (Geneva, May 9-11, 2018) was hosted by the European Alzheimer’s Disease Consortium (EADC), the Geneva Memory Clinic, the Centre Interfacultaire de Gérontologie et d’Études de Vulnerabilité (CIGEV), and Swiss Memory Clinics. Participants were European dementia experts—physicians, neurologists, (neuro)psychologists, psychiatrists, geriatricians—from non-academic and academic memory clinics, researchers from previous pertinent harmonization initiatives worldwide, methodologists, and stakeholders (https://cigev.unige.ch/files/S015/3788/2053/hnade.pdf) (see Supplemental Box S1 for institutions and their representatives and Supplemental Box S2 for individual participants and affiliations). At the workshop, presenters described the methods, results, issues, and resources from previous harmonization initiatives, and current development of tests specific to dementing neurodegenerative disorders in plenary sessions. After the plenaries, participants were assigned to specific thematic subgroups based on their expertise and leveraging published evidence, to work in parallel and find solutions to harmonize the aspects specific to their competence (neuropsychology experts: test selection; statistics experts: modeling for the generation of normative values; digital working group; potential and issues on digital-assisted testing).

All of the consensual decisions reported in Results derive from the following procedure (Figure 1).
2.1 | Thematic subgroups

Each thematic subgroup was led by one expert (two in the case of the statistical working group) to help discussions through a semi-structured approach. They elaborated on methodology (methods to define normative values; statistical approaches/modeling), issues related to neuropsychological tests (test selection; hypothesis-driven test generation), and digitally assisted testing. At the workshop, the subgroups developed proposals for defining the normal population, confounding factors and minimum sample size required to produce consistent normative values, test selection for standard assessment, and perspectives for future consideration. Decisions were proposed, discussed, refined, and ratified in subsequent plenary discussions with informal consensus procedures. For some aspects, the subgroups were tasked to further process their topics after the workshop. Subgroups completed the processing of their tasks in the following months through online meetings and provided written sections for the paper and the supplemental material. All participants were entitled to object and contribute. Divergence was settled based on published data, processed by the pertinent thematic working group after the workshop. When no further objections emerged for solutions to previous objections, we considered that the majority agreed on the latest solution.

2.2 | Consensus refining

After the workshop, the moderator (first author) incorporated the achieved consensus decisions into a first draft paper. Additional contributors were invited to provide information and knowledge based on their expertise in the field, and their comments were accommodated into the manuscript. Workshop participants and the additional contributors could access the manuscript at all phases, and proposed comments were accessible to all. Whenever objections and disagreement requiring specific expertise arose, the moderator consulted the thematic subgroups again for qualified processing. This was done through both in-person or remote meetings. Both expertise and published evidence were used to support decisions.

2.3 | Quantification of final consensus

At the end of this procedure, participants could express their (a) full agreement, (b) partial agreement (ie, agree as a first step, in view of next improvement), or (c) disagreement, and could propose reasons and further comments through a formal voting system. Those who took part in the workshop in person and the additional contributors based in Europe were invited to this final vote ($N = 47$). Because all of the proposed points came from a lengthy (although informal) consensus procedure, we set the threshold for agreement at 90%. Both the options “Full” and “Partial agreement in view of future improvement” were considered as agreement for the current decision, “Partial agreement in view of future improvement” meaning that the current solution is a required interim step toward harmonization. Answers of partial or lacking agreement required explanation of reasons. These were used to further improve the final manuscript, or processed as far as possible, fed back to all participants blinding the identity of individual responders, and stored to support next developments.

2.4 | Definition of sample size

The statistical subgroup started with a general-purpose estimation method to pinpoint the number of subjects required to compute normative values with correction parameters based on the consensual
classification of confounding factors. A simulation procedure was then used to compute the minimum sample size per language that would allow for (a) stable computations across different scenarios (eg, different distribution properties of test scores, adjustments for confounding factors) and (b) use of complex computational procedures (eg, Item Response Theory, Structural Equation Models) to allow development of flexible composite measures. Additional computational details are provided in Supplemental Section 1.

3 | RESULTS

Consensus was achieved and formalized (1) on a proposal for a Uniform Dataset analogous to that produced by the U.S. National Alzheimer’s Coordinating Center (NACC) (Table 1) for the context of use of diagnosis in memory clinics, and on enriching it with tests specific to the pathophysiology of typical Alzheimer’s disease (AD) and behavioral variant frontotemporal dementia (bvFTD); (2) on a standard definition of normal controls (Box 1), aimed to produce consistent normative values; and (3) on a standard methodology to produce harmonized normative data across tests and languages (Box 1; Table 1). The main next steps will require refining the harmonization as emerged from the discussions and to proceed toward implementation (Box 2). Forty-two of the 47 invited participants and contributors sent their final votes and comments at the final questionnaire. Formal consensus with the decisions expressed in the final manuscript was 100% for all points. Partial agreement that could not be fully accommodated in this manuscript amounted to three voters (7%) for the adoption of UDS-3 and the inclusion of FCSRT and the Story-Based Empathy Task (SET), two (5%) for the definition of normal controls, one (2%) for the inclusion and exclusion criteria of controls, none for the methodology to provide standard normative values, and two (5%) for the next harmonization steps.

3.1 | Context of use

Workshop participants agreed on the need and timeliness of a harmonization initiative for neuropsychological assessment for memory clinics and its concomitant utility for clinical research, and on the need to encourage a collaborative and representative participation of different clinical contexts and countries. Although aimed at European clinical settings, the intrinsic European cultural and linguistic heterogeneity, the high rate of migration, and the need to leverage existing resources and to align the clinical and clinical-research contexts motivated participants to consider this effort within a broader European and non-European context. This effort has been aimed primarily at clinical use, but its applicability to clinical research is straightforward.

The proposed primary objective of the anticipated standard assessment is to reliably identify MCI and progression of cognitive decline in persons referred to memory clinics, or other specialized centers, for cognitive complaints. The assessment is not designed to detect subtle deficits in the preclinical disease phase, or to grade severity of impairment at the dementia stage. Moreover, it is not designed for population screening, case finding, or finer cognitive profiling aiming at other purposes (eg, to formulate etiopathological diagnoses based on cognitive profile, or to tailor neuropsychological rehabilitation).

3.2 | Uniform data set

Workshop participants proposed and consented to adopt version 3 of the U.S. NACC Uniform Dataset (UDS-3) neuropsychological test battery, to date the most widely implemented standard battery for diagnosing MCI and measuring progression of cognitive decline in early dementia. The main reasons for partial agreement in view of future improvement in the formal voting for final agreement included the length of the proposed battery, possibly excessive for some clinics, as well as requests for even more extensive and thorough assessment; a limited added value over current harmonization in countries already using standard batteries (ie, The Netherlands and German-speaking countries); and possible issues on administration, scoring, and norms.

The following integrations were required to increase sensitivity to MCI.

3.2.1 | Adaptations

The UDS-3 largely consists of new, copyright-free versions of common neuropsychological tests that are sensitive to MCI and early dementia cognitive decline and overcome test repetition effects (Table 1). The tests were specifically developed, adapted, and normed for the elderly U.S. population. The first step for the European harmonization is to adopt European “local analogues”, that is, traditional tests with local normative values analogous to those in the UDS-3, like the WAIS digit span instead of the UDS-3 number span, or the Boston Naming Test if local norms for MINT (Multilingual Naming Test) are lacking (Table 1). Subsequent steps require adapting tests across languages and acquiring local normative data (Box 2). Another possible adaptation relates to test order. If tests are already used in local batteries, with different order due to different composition of such batteries, such order discrepancies are considered compatible with this harmonization effort at this stage.

3.2.2 | Integration

Workshop participants also agreed on including tests specific to episodic memory and emotional cognition impairment (bold in Table 1) to provide better coverage of these cognitive domains in line with Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) guidelines, and to enhance the battery’s sensitivity to the typical AD, atypical AD, and to bvFTD. Specifically, they proposed replacing the Craft Story Memory test (the UDS-3 replacement of the logical memory test) with the verbal version of the Free and Cued Selective Reminding Task (FCSR,T) and adding the Story-based Empathy Task (SET) to assess social competence. The Free and Cued Selective Reminding Task (FCSR,T) is a cued word list task providing controlled learning and retrieval conditions that enable one to...
### TABLE 1  
Uniform Dataset, based on the current US standard for research (UDS-3[1]) and the Geneva Workshop 2018 adaptation to clinical use (cUDS), mapped onto the cognitive domains and sub-functions as listed and recommended by DSM-5[2,3] for the diagnosis of neurocognitive disorders

<table>
<thead>
<tr>
<th>DSM-5-addressed domains</th>
<th>UDS-3 Test</th>
<th>Order</th>
<th>Time</th>
<th>cUDS Test</th>
<th>Order</th>
<th>Time</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>General cognitive assessment</td>
<td>MoCA</td>
<td>1</td>
<td>10'</td>
<td>MoCA</td>
<td>1</td>
<td>10'</td>
<td>Often administered at a first separate visit[5,6]</td>
</tr>
<tr>
<td>Attention</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sustained</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>selective</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>divided</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>processing speed</td>
<td>Trail making A</td>
<td>7</td>
<td>1'</td>
<td>Trail making A</td>
<td>5</td>
<td>1'</td>
<td></td>
</tr>
<tr>
<td>Executive function</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>planning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>decision making</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>working memory</td>
<td>Number span backward</td>
<td>5</td>
<td>1'</td>
<td>Digit span backward</td>
<td>3</td>
<td>1'</td>
<td>From WAIS</td>
</tr>
<tr>
<td>feedback use</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>flexibility</td>
<td>Trail making B</td>
<td>8</td>
<td>3'</td>
<td>Trail making B</td>
<td>6</td>
<td>3'</td>
<td></td>
</tr>
<tr>
<td>Perceptual-motor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>visual perception</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>constructional</td>
<td>Benson figure copy</td>
<td>3</td>
<td>5'</td>
<td>Benson figure copy</td>
<td>9</td>
<td>5'</td>
<td>or Rey Complex figure, if lacking norms</td>
</tr>
<tr>
<td>perceptual-motor praxis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gnosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social cognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>emotion recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>theory of mind</td>
<td>SET</td>
<td>7</td>
<td>10'</td>
<td>SET: Story-based Empathy task (7); <a href="https://forms.gle/muDpJLkqH6X8h9z99">https://forms.gle/muDpJLkqH6X8h9z99</a></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Learning-Memory</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>immediate recall</td>
<td>Number span forward</td>
<td>4</td>
<td>1'</td>
<td>Digit span forward</td>
<td>2</td>
<td>1'</td>
<td>From WAIS</td>
</tr>
<tr>
<td>Craft story 21–immediate</td>
<td>2</td>
<td>5'</td>
<td>FCSRT – immediate</td>
<td>4</td>
<td>10'</td>
<td>Verbal version. Normative values available in many languages (8-15)</td>
<td></td>
</tr>
<tr>
<td>short-term memory</td>
<td>Craft story 21–delayed</td>
<td>9</td>
<td>10'</td>
<td>FCSRT – delayed</td>
<td>8</td>
<td>3'</td>
<td>FCSRT</td>
</tr>
<tr>
<td>Benson figure recall</td>
<td>10</td>
<td>5'</td>
<td>Benson figure recall</td>
<td>13</td>
<td>5'</td>
<td>Category fluency and MINT have a long-term memory component</td>
<td></td>
</tr>
<tr>
<td>long-term memory</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>implicit memory</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Continues)
TABLE 1 (Continued)

<table>
<thead>
<tr>
<th>DSM-5-addressed domains</th>
<th>UDS-3 Test</th>
<th>Order</th>
<th>Time</th>
<th>cUDS Test</th>
<th>Order</th>
<th>Time</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language production</td>
<td>MINT</td>
<td>11</td>
<td>5-10'</td>
<td>MINT</td>
<td>12</td>
<td>5-10'</td>
<td>or Boston naming test, if lacking norms</td>
</tr>
<tr>
<td>Category fluency</td>
<td>6</td>
<td>3'</td>
<td></td>
<td>10</td>
<td>3'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(animals, vegetables)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Letter fluency (F, L)</td>
<td>12</td>
<td>3'</td>
<td></td>
<td>11</td>
<td>3'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>comprehension</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total time</td>
<td>52-57'</td>
<td></td>
<td></td>
<td>60-65'</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tests in italics denote “local analogues,” that is, traditional local versions of the UDS-3 tests. Bold (besides titles) denotes tests added or replaced to UDS-3 tests. “Order” denotes the presentation order in UDS-3 as from https://www.alz.washington.edu/NONMEMBER/UDS/DOCS/VER3/UDS3_npsych_worksheets_C2.pdf and the administration order of the cUDS tests as described in the Results section. Time denotes expected duration of administration estimated for patients with MCI (with CDR test score between 0.5-1) and including instructions. MINT = Multi-lingual naming test; FCSRT = Free and cued selective reminding test; verbal version(4); MoCA Montreal Cognitive Assessment.

Category fluency and MINT also have a long-term memory component.

References.

BOCCARDI ET AL.

THE JOURNAL OF THE ALZHEIMER’S ASSOCIATION

distinguish between impaired encoding (eg, due to attention disorders as in depression), storage (eg, due to hippocampal damage as in typical AD), and recall strategies (eg, due to frontal lobe dysfunction). Impaired task performance in the FCSRT correlates with in vivo AD biomarkers, even at the prodromal stage. Low total recall performance despite retrieval facilitation with cueing discriminates MCI patients subsequently converting to AD dementia with 88.6% specificity, showing better diagnostic and prognostic performance than the Logical Memory task. The FCSRT has been validated in many EU countries, with normative values available in different languages (see Table 1). The verbal version of the FCSRT has lower ceiling effects and greater dispersion of test scores, and is therefore preferred over the visual version for a standard assessment. The SET was specifically developed to assess social cognition in patients with dementia. It requires subjects to select the possible endings of stories told with cartoons, and assesses emotion and intention attribution separately, as well as causal inference as the control condition. Verbal interaction is used to ascertain proper comprehension of instructions; however, correct performance consists of selecting the correct story ending among the available pictures. Similar to the mini Socio-Emotional Assessment (mini-SEA), performance on the SET correlates with structural and functional imaging evidence of frontal lobe dysfunction and hippocampal atrophy.
BOX 1  Consensual definition of Normative Population to obtain consistent normative values across countries and tests. All of the reported criteria must be met to recruit proper harmonized control samples

Selection of normal (not super-normal) subjects

- Bias-free recruitment modalities (e.g., to guarantee representativeness for the whole country population, recruitment should avoid clustered data within just one site, and data should be gathered pairwise—for age, sex, education—within each recruitment site)
- Avoid convenience samples unless compliant with the inclusion/exclusion criteria
- Avoid voluntary exclusion of otherwise healthy individuals positive to biomarkers for brain amyloidosis, tau, or other neurodegeneration that define risk or preclinical stage for neurocognitive disorders
- Avoid voluntary exclusion of subjects with subjective cognitive decline (SCD) from samples explicitly recruited as normal controls
- Do not seek demonstration of stable cognitive health with longitudinal neuropsychological and neurological assessment ("robust norms")

Inclusion Criteria

- Age: if feasible, 40 years or older
- Self-identified as "cognitively normal"
- Denies a worrying cognitive decline
- Judged to be cognitively normal by a family member (or other knowledgeable informant); cut-off of 3.3 on the short form of the Informant Questionnaire on Cognitive Decline (IQCoDe)(1) or equivalent
- MoCA score greater than or equal to 23(2,3) (MMSE greater than or equal to 27 if MoCA cannot be used). Please note: values to be corrected by age and education

Exclusion Criteria

Clinical findings

- Sensory or motor deficits interfering with test administration
- Continuous moderate-to-intense pain
- Current psychiatric diagnosis (including major depression). Geriatric Depression Scale (15-items) score of 6 or greater (4)

Medical history

- Head injury with loss of consciousness for more than 5 minutes
- General anesthesia within the last 3 months
- Prior recurrent psychiatric disorder requiring hospitalization
- Use of psychoactive drugs, alcohol abuse
- Significant cerebrovascular disease (e.g., TIA, stroke, general atherosclerosis)
- Severe active medical condition (cancer, organ failure, unstable heart condition) that may interfere with test administration

Convenience samples

- May be used if compliant with the above features
- May be used ad interim when proper samples are unavailable
- Specific research samples (only SCD, or composed of subjects all having specific risk factors) are not appropriate

References.


3.2.3 Administration procedure

The total administration time for the CUDS is estimated to be 60-65 minutes for patients with MCI. The delay interval required between immediate and delayed trials of the FCSRT12,14 is filled with administration of the non-verbal Trail Making A and B and SET13, allowing the delay to be free from interference from other verbal tasks27 (Table 1).

3.3 Harmonized methodology to produce normative values

Workshop participants highlighted the priority of defining normative values based on standard methodology to guarantee consistent assessment of performance across tests and languages. Therefore, they
developed a standard definition of the normal population, defined how to code confounders of normative data consistently, and proposed the minimum necessary requirements for the normative sample size.

### 3.3.1 Normal population

The normal population that will provide harmonized normative data for the proposed context of use must consist of individuals without cognitive decline, functional impairment due to cognitive deterioration, or major clinical conditions that could interfere with cognition (Box 1). Main reasons for partial agreement in view of future improvement in the formal voting for final agreement relate to the potential appropriateness of robust controls. At present, recruitment of normal controls from the general population should not exclude cognitively unimpaired individuals who may have positive biomarkers or specific risk factors (e.g., APOE ε4 allele) for AD, nor should it target the “super-normal” individuals (also described as “robust controls”) with longitudinal evidence of preserved cognition, although this will plausibly be a required future development (Box 2). Convenience samples of cognitively normal individuals may be employed only if they were recruited to serve as normal controls and if compliant with the criteria reported in Box 1. Convenience cognitively normal samples should not consist of individuals recruited as target experimental cases (e.g., people with subjective cognitive decline or otherwise at risk for neurocognitive disorders). One reason for partial agreement in view of future improvement in the formal voting for final agreement on the individual exclusion/inclusion criteria in Box 1 consisted in disagreement on excluding people who experienced post-traumatic loss of consciousness without significant memory loss.

### 3.3.2 Demographic factors affecting neuropsychological performance

Standard classification was proposed for age, sex, and education, the main variables that often affect neuropsychological test performance and should be accounted for in normative data. Panelists agreed that normative data and test cut-off values should be stratified across 6 age decades, from the 40 to the upper 90+ year age categories, with balanced sex and education level in each age category. To code education consistently across countries, we adopted the three-level coding system from the International Standard Classification of Education (ISCED) of the United Nations Educational, Scientific and Cultural Organization (UNESCO) [https://ec.europa.eu/education/international-standard-classification-of-education-isced_en](https://ec.europa.eu/education/international-standard-classification-of-education-isced_en). This
system includes (1) compulsory (primary and secondary) education (age range: min = 5, max = 17), (2) upper-secondary education (age range for EU countries: min = 12, max = 19), and (3) post-secondary education (age range: min = 18, max ≥ 22) (Supplemental Table S3).

3.3.3 Sample size

The above classification scheme defines 36 cells that encompass six age classes by three educational levels by two sexes (Supplemental Table S4). Our sample size computation estimated that 10 subjects per cell would suffice to perform general linear models (see Supplemental Section-1 for computational details). Data from this carefully constructed sample can be used to normalize standard scores or modern psychometric scores derived from Item Response Theory or Structural Equation Modeling. An estimated sample size of N = 330 subjects per language provides a fair trade-off between feasibility and reliability, whereas taking into account the stratifications for age, education, and sex (although cells covering rare populations, eg, age 90+ with high education, may be hard to fill) (Supplemental Table S4).

3.3.4 Test validation

The cUDS neuropsychological tests should ideally be copyright-free versions, analogous to the U.S. UDS-3, which are adapted to the various European target languages. They should be validated in all European languages, taking care to include linguistic minorities in validation studies.

Achieving a standard assessment of this kind requires additional research steps (see roadmap outlined in Box 2). A reason for partial agreement in view of future improvement in the formal voting for final agreement on such roadmap consisted in the difficulty of harmonizing and producing culture-free standardized social cognition among countries. Another main hurdle lies in the fact that some countries already use CERAD or other batteries as local standards. However, using the cUDS, as a common set of UDS-3 or analogous tests with locally appropriate normative data was seen as a first practical step towards harmonization.

4 DISCUSSION

This work defines the first steps toward a standard assessment of people with cognitive complaints attending European memory clinics or participating in clinical research on MCI in Europe. This assessment includes a standard neuropsychological test battery with a harmonized methodology to produce normative data and cut-off values for impairment. Besides leveraging on previous harmonization efforts, the bidirectional collaboration with clinicians from specialized non-academic centers is a new and important step necessary to fit the needs and constraints of both clinical practice and research. With evolving clinical criteria and the availability of biomarkers with specific diagnostic value, it is now even more important to assess patients consistently with a precise definition of the target disorder, with the demonstrated informative value of diagnostic biomarkers, and across centers. With standard assessment, patients could get second opinions or receive follow-up examinations in different centers without the need to repeat existing baseline assessments. Costs, as well as practice effects, would be reduced; benefits for data pooling and comparability of studies in clinical research are straightforward. The reliability of diagnostic procedures for MCI may improve and approach the quality standards of other clinical conditions.

The proposed standard assessment is not designed to ascertain the pathophysiology underlying cognitive impairment, but to detect the presence and possibly progression of objective cognitive decline that may be due to a neurodegenerative condition. Further exploration, increasingly performed through biomarkers, is required to formulate an etiopathological diagnosis or to identify specific clinical needs.

The cUDS is based on the U.S. UDS-3. Extensive work was undertaken to develop the UDS-3 for the cognitive assessment of individuals with MCI, before it was adopted as a standard cognitive assessment in all federally funded U.S. AD research centers. Although designed to improve CERAD-NAB performance, to date there is no systematic evidence on the ability of UDS-3 in detecting MCI. The cUDS is expected to outperform the CERAD-NAB in the detection of MCI. Although overlapping with CERAD-NAB and its “plus” version (Trail Making, Figures copy/recall; Boston Naming Test and Verbal Fluency tasks), the Montreal Cognitive Assessment (MoCA), included in the cUDS, was shown to clearly outperform the CERAD’s Mini-Mental State Examination. The CERAD word list was demonstrated to be less sensitive than the California Verbal Learning test, and its diagnostic and prognostic performance for AD-MCI was demonstrated to increase by adding the FCSRT (32). In addition, the SET allows early detection of impairment also in patients affected by frontotemporal lobar degeneration, who may perform well on typical cognitive tests. Overall, cUDS overlaps with CERAD to a considerable extent but is devised to be more sensitive to mild, atypical, and non-AD conditions. Although CERAD-plus has specific normative values for US-English and German populations (Supplemental Table S2), UDS-3 local analogues are frequently used in European memory clinics with local norms for most countries. Alignment with the NACC UDS-3 sets this European proposal up for improved sensitivity to MCI due to different neurodegenerative causes and for a possibly wider international consensus.

In addition to CERAD and UDS-3, the Neuropsychological Normative Project (CN-NORM) (Supplemental Table S2) has selected and recommended tests for harmonized assessment of MCI for Chinese memory clinics. Besides some overlap with the cUDS (eg, trail making, digit span and fluency tests, and the memory binding test that, like the FCSRT, uses a controlled learning paradigm minimizing the use of individual strategies), the CN-NORM battery covers cognitive domains more extensively than the cUDS. In particular, it allows a more thorough assessment of attentional, perceptual-motor, and social functions. This main difference between the cUDS and the CN-NORM consists of the fact that the cUDS aims to provide objective evidence of impairment without the aim to identify pathophysiology, because it is devised for a biomarker-based diagnostic procedure. This differs from
the CN-NORM initiative, that is not expressly restricted to biomarker-based procedures and consistently provides more thorough cognitive assessment requiring additional measures. A short battery, however, has greater potential to be adopted in EU countries. Nevertheless, future developments aimed to improve the assessment of cognitive and social functions not assessed by the current cUDS may consider the Chinese standard and seek further consistency across Western and Eastern countries.

Previous initiatives selecting and recommending tests for standard assessment relate to research on preclinical AD (European Prevention of Alzheimer’s Disease [EPAD]) and Alzheimer’s disease cognitive composite [PACC], Supplemental Table S2,36,37 a different context of use relative to the cUDS. The brevity of PACC makes it an interesting option for detecting MCI, and decreased scores were associated with MCI in research cohorts.37,38 Its potential in detecting MCI patients in memory clinics, not yet explored, is expected to be lower than cUDS. First, cUDS also includes a test for MCI due to non- or atypical AD. Moreover, the mentioned limitations of logical memory test16,17,19–22 and of the Mini-Mental State Examination (MMSE)22 are consistent with the demonstrated improved performance of PACC after removing MMSE and including semantic and executive assessment.39,40 A PACC-like composite can be derived from the cUDS; however, future developments of initiatives aimed at assessing individuals at preclinical (PACC, EPAD) and clinical (cUDS, UDS-3, CN-NORM) disease stage may try to seek consistency across each other (eg, FCSRT is already included both in PACC and cUDS). Bridging these different contexts of use may provide continuity of assessment, thereby sparing money and effort, if preclinical assessment should be adopted for future population screening in the future.

The methodology we propose to produce normative data and cut-off values for impairment is not new in terms of the need for correction itself but is new as it tries to align the validation of different tests and of tests in different languages to a common methodological standard. The age range of controls providing normative values (40+) allows the detection of early onset cases; the standard classification of education, although based on only three levels, guarantees a reliable comparison across countries characterized by very different educational systems, which can hardly be captured in a harmonized framework. Although such compromises are required to achieve a minimum and feasible harmonization, neuropsychological research groups are encouraged to provide additional, finer normative values for research aims, for other contexts of use, or to better account for less represented groups (eg, elderly with very low educational attainments). Further stratification for variables like residence in urban versus rural areas may also be included if possible. Other possibly confounding variables, like ethnicity, may have a less consistent effect in Europe than in the United States, as migration waves are currently variable and variably handled. On the other hand, norms based on "robust" control samples (ie, subjects whose normal cognition is documented over longitudinal evaluations), or on individuals free from pathology or risk factors for neurocognitive disorders, may also be of interest for research aims, and may be required in future clinical applications, but are not pertinent to the proposed standard assessment aimed at detecting mild impairment from a clinical rather than pathophysiological point of view.

Further fine-tuning of the cUDS is warranted to meet practical constraints to implementation, to assess other dimensions (eg, motivational level, malingering), or to provide more balanced testing of the different cognitive domains, similar to the CN-NORM battery. Newly developed hypothesis-driven tests may increase sensitivity (Supplemental Section 2). Moreover, consistency of administration should also be achieved across centers, raters, and time through specific methodological procedures defining standard administration, data entry, score computation, and ascertaining compliance and reliability over time, similar to other diagnostic procedures. Score conversion tools and demonstration of backward compatibility with local batteries are also required, to shift to the new standard, as is the development of copyright-free tests analogous to UDS-3, of local normative data (Box 2), and of digital infrastructure to assist the harmonized assessment (Supplemental Section 3).

Within this initiative, we have tried to take advantage of the knowledge and experience gained in different research and clinical contexts. Although not entirely new in nature, this is among the first efforts trying to converge such knowledge into a comprehensive harmonized procedure for the reliable assessment of patients with MCI possibly serving both clinical and research aims. This approach was maximized by the support and participation of several relevant consortia (Supplemental Box S1), most importantly the European Alzheimer’s Disease Consortium. The Alzheimer’s Association International Society to Advance Alzheimer’s Research and Treatment (ISTAART) Professional Interest Area on Cognition has initiated a Workgroup on Harmonization of Assessment that will also help bring the relevant groups together to review, evaluate, and make recommendations on convergence. It should be noted, however, that representation in this process was limited by logistics and feasibility, penalizing clinicians not engaged in research at this step. Another major limitation that we did not address consists of the heterogeneous definition of MCI. Different definitions have been applied, but ideally a consensus should be achieved upstream of a test battery selection. Nevertheless, a standard selection of tests is a valuable starting point for subsequent more thorough harmonization. The opportunity to benefit from digital technology (Supplemental Section 3) to assist the standardized assessment exists but requires additional consensual development and implementation while respecting clinical and scientific principles underlying standard assessment for neurocognitive disorders. Future developments may allow for more accurate and reliable assessment, especially relative to confounders that cannot be stratified and corrected for in a satisfactory way based on pen-and-pencil tests and traditionally collected normative values. Finally, the number and complexity of the variables processed in this consensus was so high that formal traditional consensus procedures (eg, Delphi panel) could not be applied from the beginning to each of them for reasons of feasibility. Future developments should try to use formal methods. Despite these limitations, this first step toward harmonization in memory clinics can help develop more modern and efficient clinical procedures for neurocognitive disorders through the consistent definition of normal controls and
methodological procedures for the production of norms, the harmonization across tests and languages, and the attempt to seek for extra-European convergence. It will optimize costs and reliability, aligning diagnostic procedures across centers, with the demonstrated informative value of diagnostic tools, and with the therapeutic value of available treatment expected to slow down progression and improve the quality of life of patients and caregivers41 (34).

ACKNOWLEDGEMENTS
The workshop was organized within the Laboratoire de Neuroimagerie du Vieillissement and the Memory Clinic at Geneva University and University Hospitals (Giovanni Frisoni) and hosted by the European Alzheimer’s Disease Consortium (EADC; Bengt Winblad, Lutz Frölich) and the Centre Interfacultaire de Gérontologie et Études des Vulnérabilités et Vieillissement (CIGEV; Matthias Kliegel; University of Geneva). Balanced participation of academic and non-academic clinicians from the different linguistic Swiss regions was allowed by Swiss Memory Clinics (president: Markus Bürge). Representatives of stakeholders were Bengt Winblad and Lutz Frölich for the EADC, Jean Georges for Alzheimer’s Europe, Andreas U. Monsch for Swiss Memory Clinics, Alzheimer Forum Switzerland and Swiss Association of Neuropsychologists, Matthias Kliegel for Age-NT, and Stefano Cappa for the European Federation of Neuropsychological Societies. ISTAART PIA-Cognition and EADC supported follow-up coordination meetings. The Geneva Workshop was funded by the Swiss National Science Foundation (grant number ISZEZO_180198), Alzheimer Forum Switzerland and MindMaze. Part of the follow-up work was supported by LANVIE (Laboratoire de Neuroimagerie du Vieillissement - Giovanni Frisoni) and the Laboratory of Cognitive Aging (Matthias Kliegel) at University of Geneva. M. Boccardi was supported in part by the EU-EFPIA Innovative Medicines Initiative 2 Joint Undertaking (grant no. 115952; recipient: Giovanni Frisoni, AMYPAD project). I. Bos and P.J. Visser were supported by the Innovative Medicines Initiative Joint Undertaking. P.J.V. participation was also supported by ZONmW. The participation of W. Kukull was supported by NIH/NIA (NIH grant no. U01 AG016976); participation of P. Sachdev was supported by NHMRC Australia. C. Ferrari was supported by the Italian Ministry of Health (Ricerca Corrente).

CONFLICT OF INTEREST
I. Bos received grants from Innovative Medicine Initiative Joint Undertaking. H. Brodaty received personal fees from Nutrivia Australia. S. Cappa received personal fees from Biogen. W. Kukull received grants from National Institutes of Health/National Institute on Aging (NIH/NIA). F. Oliveira received personal fees from Gerson Lehrman Group and Atheneum Partners.

O. Peters received personal fees from Biogen; grants and personal fees from Roche; and grants from Novartis, Janssen, Pharmatrophix, and Eisai.

P. Sachdev received grants from NHMRC Australia and personal fees from Biogen Pharmaceuticals.

D. Salmon received consulting fees from Biogen, Inc. and Aptinyx, Inc.

Teipel received personal fees from Roche Pharma AG, MSD Sharp & Dohme GmbH, and Biogen.

P.J. Visser received grants from Innovative Medicine Initiative, ZONmW, and Biogen.

B. Winblad received advisory board fees from AlzeCure, Alzilnova, and Axon Neuroscience.

The other authors report no conflict of interest.

ORCID
Marina Boccardi https://orcid.org/0000-0001-6744-8246

REFERENCES
1. APA. Diagnostic and Statistical Manual of Mental Disorders. 5th. 2013.


APPENDIX: COLLABORATORS

SUPPORTING INFORMATION
Additional supporting information may be found online in the Supporting Information section at the end of the article.