Abstract

Background: The identification of the gene mutation causing Huntington disease has raised hopes for new treatments to ease symptoms and slow functional decline. As such, there has been a push towards designing efficient pharmacological trials (i.e., drug trials), especially with regard to selecting outcomes measures that are both brief and sensitive to changes across the course of the disease, from subtle prodromal changes, to more severe end-stage changes.

Objectives: Recently, to aid in efficient development of new HD research studies, the National Institute of Neurological Disorders and Stroke (NINDS) published recommendations for measurement selection in HD. While these recommendations are helpful, many of the recommended measures have little published data in HD. As such, we conducted a systematic review of the literature to identify the most common outcomes measures used in HD clinical trials.

Methods: Major medical databases, including PubMed, Embase, CINAHL, and the Cochrane Central Register of Controlled Trials, were used to identify peer-reviewed journal articles in English from 2001 through April 2013; 151 pharmacological trials were identified.

Results: The majority of HD clinical trials employed clinician-reported outcomes measures (93%); patient reported outcome measures (11%) and observer reported outcome measures (3%) were used with much less frequency.

Conclusions: We provide a review of the most commonly used measures across these trials, compare these measures to the clinical recommendations made by the NINDS working groups, and provide recommendations for selecting measures for future clinical trials that meet the Food and Drug Administration standards.

Keywords: Huntington disease, clinical trials, outcomes measures

Huntington disease (HD) is an autosomal dominant neurodegenerative disease affecting approximately 3 per 100,000 people worldwide [1]. Individuals with the abnormal CAG expansion in the HTT gene experience a multitude of behavioral, cognitive and motor symptoms over the course of the disease. Such symptoms include depression, anxiety, personality change, irritability, dementia, chorea, imbalance, clumsiness, falls, and swallowing difficulty [2]. Symptoms often begin insidiously around age 40, and progress steadily over 15–20 years, leading to death [3].

The identification of the gene mutation causing HD, as well as continued advances in understanding the pathogenesis of the disease, has raised hopes for evaluating new clinical compounds designed to alleviate symptoms and to slow functional decline, if not to cure the disease outright. However, most outcomes
measures currently used in HD pharmacological trials are narrow in focus and do not adequately capture changes in function at the most meaningful level, for patients [4–6]. Without sensitive and valid outcomes measures, it is impossible to evaluate the effectiveness of potential treatment interventions.

In addition, the evaluation of new clinical compounds is complicated by other factors. HD is considered a rare or orphan disease, making very large clinical trials impractical. Furthermore, it is currently impossible to ascertain, for at-risk individuals nearing the typical age of onset, when the disease might “manifest” itself [7]. Thus, if a 38-year old is given a compound and does not exhibit symptoms for several more years, there is no way to know whether the compound prevented or delayed clinical onset for that individual. Given such complexities, well-designed pharmacological trials are paramount in ultimately deciding whether a new treatment is effective in HD. One of the best ways to make these clinical trials more efficient and robust is to maximize the sensitivity of our clinical assessment tools.

The HD community has long recognized the need for more sensitive, HD-specific measurement. Most recently, the National Institute for Neurological Disorder and Stroke (NINDS) has established a group of HD clinicians and researchers to make recommendations for common data elements (CDE) in HD clinical trials and research. The NINDS CDE groups included working groups in: Motor, Imaging and Biomarkers, Biochemical Markers, Genetics, Epidemiology/Environment, Function Outcomes/PROs, Behavior/Psychiatry, Pathology, Operations, Cognition, and Scale Metrics and Statistics [8]. Each group was to review the state of the science to determine which clinical measures are the most useful/sensitive in an HD population. To this end, they have published a list of recommendations for measurement selection in clinical research in HD (see Table 1 for a summary of recommended measures) [8]. The utility of each measure was classified as follows: core (recommended for use in all HD studies), supplemental (recommended for targeted use in HD studies), or exploratory (not enough data to make a full recommendation, but some evidence to suggest utility in HD). While these recommendations provide an excellent starting point for measurement selection for clinical studies, each working group developed their own criteria for determining classifications in a manner not consistent across working groups. For example, some groups decided to recommend measures based on expert opinion, some based on literature review (of varying levels of rigor).

The need for sensitive assessments in HD research is particularly acute in pharmacological trials (i.e., drug trials). Specifically, in determining the efficacy of a drug, the Food and Drug Administration (FDA) requires evidence to support dosing selection, safety, tolerability, and the ability of the new drug to improve some specific facet(s) of the disease in question (and when applicable, compared to other existing treatments) [9]. Clinical trials might use a variety of primary and secondary outcomes measures which can include clinical outcomes assessments (COAs), biomarkers and animal models. Specifically, COAs can fall into one of three categories: patient-reported outcome (PRO) assessment, observer reported outcome assessment (ObsRO), or clinical-reported outcome (ClinRO) assessment. PROs are self-report measures (i.e., responses come directly from the patient without any interpretation by a clinician/observer) that are focused on evaluating health [10]. ObsRO are measures that are made by an individual that knows the patient, but does not necessarily have professional training (e.g., family member, friend), and ClinRO are assessments that are made by physicians using clinical judgment and/or interpretation. While a PRO can evaluate all aspects of health (direct assessment of symptoms, observable and unobservable behaviors), ObsRO and ClinRO assessments can only be used to evaluate observable behaviors (which does not include the direct assessment of symptoms or unobservable behaviors and feelings). Qualifying assessments must be standardized (administration and responses), have acceptable psychometric properties including validity data in the targeted patient population.

Given the relative importance of maximizing HD pharmacological trials, the recognition in the field that sensitive HD-specific measures are lacking, and the fact that hundreds of trials have already been conducted in HD, the purpose of this paper was to systematically review the literature to report the most commonly used measures in previous HD clinical trials. We provide a summary of the most frequently used measures identified by the systematic review, compare and contrast these measures with those recommended by the NINDS CDE working groups, and provide recommendations for future measurement selection based on this information.

METHODS

We completed a comprehensive search of major medical databases including PubMed, Embase, Cumu-
Table 1
Specific outcomes measures recommended by the NINDS HD CDE working groups

<table>
<thead>
<tr>
<th>Outcomes measure</th>
<th>NINDS classification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CLINICIAN REPORTED OUTCOMES (ClinRO)</strong></td>
<td></td>
</tr>
<tr>
<td>Motor Functioning</td>
<td></td>
</tr>
<tr>
<td>NIH Toolbox Motor Function Endurance (2-min. walk test);</td>
<td>Exploratory</td>
</tr>
<tr>
<td>Locomotion (4-meter walk test)</td>
<td>Exploratory</td>
</tr>
<tr>
<td>Timed Up and Go [57]</td>
<td>Exploratory</td>
</tr>
<tr>
<td>UHDRS Motor Exam [18]</td>
<td>Core</td>
</tr>
<tr>
<td>10-Meter Walking Test [58]</td>
<td>Exploratory</td>
</tr>
<tr>
<td>Cognitive Functioning</td>
<td></td>
</tr>
<tr>
<td>Circle Tracing [59]</td>
<td>Supplemental</td>
</tr>
<tr>
<td>Cued Movement Sequencing [60]</td>
<td>Supplemental</td>
</tr>
<tr>
<td>Hopkins Verbal Learning Test-Revised (HVLT-R) [62]</td>
<td>Supplemental</td>
</tr>
<tr>
<td>Map Search Task [63]</td>
<td>Exploratory</td>
</tr>
<tr>
<td>Mental Rotation [64]</td>
<td>Exploratory</td>
</tr>
<tr>
<td>Montreal Cognitive Assessment (MoCA) [65]</td>
<td>Exploratory</td>
</tr>
<tr>
<td>Phonemic Verbal Fluency (PVF) [60]</td>
<td>Supplemental</td>
</tr>
<tr>
<td>Self-Paced Tapping [66]</td>
<td>Core</td>
</tr>
<tr>
<td>Simple and Two-Choice Reaction Time [60]</td>
<td>Supplemental</td>
</tr>
<tr>
<td>Spot the Change [31]</td>
<td>Supplemental</td>
</tr>
<tr>
<td>Stroop Color Naming [23]</td>
<td>Core</td>
</tr>
<tr>
<td>Stroop Interference [23]</td>
<td>Supplemental</td>
</tr>
<tr>
<td>Symbol Digit Modalities Test [24]</td>
<td>Core</td>
</tr>
<tr>
<td>Trailmaking Test [67]</td>
<td></td>
</tr>
<tr>
<td>Verbal Fluency [22]</td>
<td></td>
</tr>
<tr>
<td><strong>Supplemental</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Emotional/Behavioral Functioning</strong></td>
<td></td>
</tr>
<tr>
<td>Apathy Evaluation Scale [68]</td>
<td>Supplemental</td>
</tr>
<tr>
<td>Apathy Scale [69]</td>
<td></td>
</tr>
<tr>
<td>Columbia Suicide Severity Scale [70]</td>
<td></td>
</tr>
<tr>
<td>Concise Health Risk Tracking Scale [71]</td>
<td>Supplemental</td>
</tr>
<tr>
<td>Hospital Anxiety and Depression Scale [44]</td>
<td></td>
</tr>
<tr>
<td>Intimacy Scale [12]</td>
<td></td>
</tr>
<tr>
<td>Problem Behaviors Assessment-Short [35]</td>
<td></td>
</tr>
<tr>
<td><strong>Functional Limitations</strong></td>
<td></td>
</tr>
<tr>
<td>Physical Performance Test (PPT) [72]</td>
<td>Exploratory</td>
</tr>
<tr>
<td>UHDRS Functional Assessment Checklist [9]</td>
<td>Core (for dx only)</td>
</tr>
<tr>
<td>UHDRS Independence Scale [9]</td>
<td>Core (for dx only)</td>
</tr>
<tr>
<td>UHDRS Total Functional Capacity [9]</td>
<td>Core (for dx only)</td>
</tr>
<tr>
<td><strong>Other Assessments</strong></td>
<td></td>
</tr>
<tr>
<td>OSU TBI Form [73]</td>
<td>Supplemental</td>
</tr>
<tr>
<td>The Retrospective Lifestyle Questionnaire [74]</td>
<td>Supplemental</td>
</tr>
<tr>
<td><strong>PATIENT REPORTED OUTCOMES (PROs)</strong></td>
<td></td>
</tr>
<tr>
<td>Cognitive Functioning</td>
<td></td>
</tr>
<tr>
<td>Florida Obsessive Compulsive Inventory [78]</td>
<td>Supplemental</td>
</tr>
<tr>
<td>Lifetime Cognitive Activity [76]</td>
<td></td>
</tr>
<tr>
<td>Passa Inventory-OCD-Week-U-Revised [77]</td>
<td>Supplemental</td>
</tr>
<tr>
<td>Emotional/Behavioural Functioning</td>
<td></td>
</tr>
<tr>
<td>Concise Health Risk Tracking Scale [71]</td>
<td></td>
</tr>
<tr>
<td><strong>Functional Limitations</strong></td>
<td></td>
</tr>
<tr>
<td>FURST/CHDI [78]</td>
<td>Exploratory</td>
</tr>
<tr>
<td>HD Work Function Scale [79]</td>
<td>Exploratory</td>
</tr>
<tr>
<td>Quality of Life</td>
<td></td>
</tr>
<tr>
<td>EuroQol 5-D (EQ-SD) [80]</td>
<td>Exploratory</td>
</tr>
<tr>
<td>Huntington's Disease health-related Quality of Life questionnaire (HDQoL) [50]</td>
<td>Supplemental</td>
</tr>
<tr>
<td>NeuroQoL [48]</td>
<td>Exploratory</td>
</tr>
<tr>
<td>PROMIS [46]</td>
<td>Exploratory</td>
</tr>
<tr>
<td>SF-36 [81]</td>
<td>Exploratory</td>
</tr>
</tbody>
</table>
Table 1 (Continued)

<table>
<thead>
<tr>
<th>Outcomes measure</th>
<th>NINDS classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sickness Impact Profile (SIP) [81]</td>
<td>Exploratory</td>
</tr>
<tr>
<td>World Health Organization Disability Assessment Schedule (WHODAS) [82]</td>
<td>Exploratory</td>
</tr>
<tr>
<td>World Health Organization Disability Assessment Schedule 2.0 (WHODAS 2.0) [83]</td>
<td>Exploratory</td>
</tr>
<tr>
<td>Other Assessments</td>
<td></td>
</tr>
<tr>
<td>Cambridge Brain Repair Centre (BRC) HD Sleep Questionnaire [84]</td>
<td>Supplemental</td>
</tr>
<tr>
<td>Food Frequency Questionnaire [85]</td>
<td>Supplemental</td>
</tr>
<tr>
<td>Godin Leisure-Time Exercise Questionnaire [86]</td>
<td>Supplemental</td>
</tr>
<tr>
<td>NIH Diet History Questionnaire (DHQ) [87]</td>
<td>Supplemental</td>
</tr>
<tr>
<td>Nurse’s Health Study Questionnaire (NHQ) [88]</td>
<td>Supplemental</td>
</tr>
<tr>
<td>PD DOC Mini Environmental Risk Questionnaire</td>
<td>Supplemental</td>
</tr>
<tr>
<td>Pittsburgh Sleep Quality Index [89]</td>
<td>Supplemental</td>
</tr>
<tr>
<td>Scale for Outcomes of Parkinson’s disease-Sleep [90]</td>
<td>Supplemental</td>
</tr>
</tbody>
</table>

In Table 1, we utilize the systematic review to identify the most frequently utilized outcomes measures in HD clinical research. For the most frequently used measures, we also reviewed available psychometric data in HD and/or other clinical populations. This data, in conjunction with the recommendations made by the NINDS, was used to make recommendations for measurement selection for future HD clinical trials. Psychometric data in HD (especially evidence of responsiveness to change in HD), was weighted more heavily than data in other clinical populations for making recommendations. Furthermore, in cases where multiple measures might have good evidence for inclusion, we often selected the measure with the most evidence to support its utility in HD.

RESULTS

The combined searches yielded 1060 abstracts. We eliminated 245 duplicate records, 18 non-English or Spanish articles, 206 review articles, 48 that did not include original data, 36 non-HD articles, 21 animal studies, 12 others, and 323 studies that were either non-pharmacological interventions or observational studies; a total of 151 articles were retained for this review (see Fig. 1). Table 2 provides a breakdown of the different types of primary, secondary and exploratory outcomes measures used in the HD pharmacological studies using FDA COA categories: specifically, most studies used clinician rated measures (i.e., 93%), whereas only a small number of clinical trials used PROs (11%) or ObsRO’s (3%). Table 3 highlights the specific outcome measures, by COA category, used in HD clinical trials. ClinRO measures included assessments of motor function, functional limitation, emotional/behavioral function, and cognitive function. The most frequently used ClinRO motor measures included the Unified Huntington’s Disease Rating Scale (UHDRS) motor exam, Abnormal Involuntary Movements Scale (AIMS), Marsden & Quinn Chorea Severity Scale, and the Quantitative Neurological Examination (QNE), whereas the most frequently used ClinRO functional limitation measures included the Total Functional

Table 2
Records reviewed for eligibility (n=471)

- Records excluded (n=523)
  - Reviews (n=468)
  - Non-Original Data (n=48)
  - Non-HD (n=36)
  - Animals (n=21)
  - Other (n=12)

Pharmacological Drug Trials included in analysis (n=151)

Records identified through PubMed, Embase, CINAHL, CCRCT (n=1060)

Records excluded (n=263)

Duplicates (n=245)

Non-English/-Spanish (n=18)

Records Reviewed (n=797)

Records reviewed for eligibility (n=474)

Pharmacological Drug Trials included in analysis (n=151)

Fig. 1. Flow diagram of the studies included in this systematic review.

Capacity Scale, the Functional Assessment Scale and the Independence Scale from the UHDRS. In addition, the most frequently used emotional/behavioral measures were the UHDRS Behavioral Exam and the Hamilton Rating Scale for Depression (HAM-D), and the most frequently used ClinRO measures of cognition were from the UHDRS (Verbal Fluency, Symbol Digit Modalities Test and the Stroop). The Clinical Global Impression Scale was also used with some frequency as a ClinRO measure. PROs and ObsRO were used infrequently.

Of the measures that were identified in the systematic review, only 9 were included in the recommendations from the NINDS (see Table 4). Eight of these were ClinRO measures: a motor functioning measure (the UHDRS Motor Exam), several cognitive measures (Verbal Fluency, Stroop, SDMT, and Trailmaking), and several Functional Assessment measures (UHDRS TFC, UHDRS Functional Assessment Scale and UHDRS Independence Scale). Our recommendations for measurement selection for future pharmacological trials are also included in Table 4.

DISCUSSION

Results from the systematic review indicated that most HD pharmacological trials use ClinRO measures as their primary endpoints. Furthermore, there is a surprising lack of PROs and ObsRO measures in these studies. In addition, there is only limited overlap between the measures that have been recommended by the NINDS, and the measures that have hitherto been used in HD clinical trials. Below, we discuss the strengths and weaknesses of the measures that have thus far been used in HD clinical trials, compare these measures to the NINDS CDE recommendations, discuss our recommendations for future measurement selection, and highlight areas where additional work is needed.

ClinRO measurements (used in 93% of HD clinical trials)

The majority of the outcomes measures used in HD pharmacological trials are clinician-rated (93% of clinical trials in HD employed ClinRO measures). This is not particularly surprising given the paucity of HD-specific PRO and ObsRO measures, as well as the concerns about the reliability of self-report data, especially in individuals with later stage HD [11–17]. ClinRO measures include assessments of motor functioning, cognitive functioning, functional limitations, and emotional/behavioral functioning.

ClinRO motor functioning measures

Among the more commonly used motor functioning measures are the UHDRS Motor Exam [18], Abnormal Involuntary Movements Scale (AIMS) [19], Marsden and Quinn Chorea Severity Score [20], and Quantitative Neurological Exam (QNE) [21] (see Table 3). The

Table 2

<table>
<thead>
<tr>
<th>Types of clinical outcome assessments</th>
<th>Number of articles (out of 151)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient Reported Outcomes (PRO)</td>
<td>15 (11%)</td>
</tr>
<tr>
<td>Observer Reported Outcomes (ObsRO)</td>
<td>4 (3%)</td>
</tr>
<tr>
<td>Clinical Reported Outcomes (ClinRO)</td>
<td>140 (93%)</td>
</tr>
<tr>
<td>Other (biomarkers, MRI, physiology)</td>
<td>46 (30%)</td>
</tr>
</tbody>
</table>

Table 3

<table>
<thead>
<tr>
<th>Types of clinical outcome assessments</th>
<th>Number of articles (out of 151)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient Reported Outcomes (PRO)</td>
<td>15 (11%)</td>
</tr>
<tr>
<td>Observer Reported Outcomes (ObsRO)</td>
<td>4 (3%)</td>
</tr>
<tr>
<td>Clinical Reported Outcomes (ClinRO)</td>
<td>140 (93%)</td>
</tr>
<tr>
<td>Other (biomarkers, MRI, physiology)</td>
<td>46 (30%)</td>
</tr>
</tbody>
</table>
### CLINICIAN REPORTED OUTCOMES (ClinRO)

**Motor Functioning**
- UHDRS Motor Exam [18]: 61
- Abnormal Involuntary Movement Scale (AIMS) [19]: 17
- Marsden & Quinn Chorea Severity Scale [20]: 6
- HD Motor Rating Scale [91]; Rockland-Simpson Dyskinesia Rating Scale [92]: 1

**Cognitive Functioning**
- UHDRS Cognitive Exam [18]: 46
- Symbol Digit Modalities Test [24]: 38
- Mini-Mental State Exam [93]: 21
- Trailmaking Test [67]: 9
- WAIS Digit Span [94, 95]: 5
- Benton Visual Retention Test [96]; Buschke Selective Reminding Test [97]: 4
- AD Assessment Scale-cognitive [98]; Raven's Progressive Matrices Test [99]; WAIS Digit Symbol [94, 95]: 3
- Hopkins Verbal Learning Test [100]; Wechsler Memory Scales [101, 102]; WAIS Digit Symbol [94, 95]; WAIS Block Design [94, 95]; WAIS FSIQ [94, 95]; RBANS [103];
- Brief Test of Attention [104]; CVLT [105]; CANTAB [106]; CERAD Verbal Learning Test [107]; Design Fluency Test [108]; Dichotomous Listening Test [109]; Digit Ordering Test [110];
- Go/No Go Test [111]; Kohs Cubes Test [112]; Luria Nebraska mental rotation item [113];
- PPVT [114]; Recurring Figures Test [115]; Road Map Test [116]; Ruff Fugl-Meyer Fluency Test [117]; Synnott Kest Test [118]; Visual Form Discrimination [119]; Wisconsin Square Picture Memory Test [120]; WAIS Letter Number Sequencing [94, 95]: 1

**Functional Limitations**
- UHDRS Total Functional Capacity Scale [18]: 48
- UHDRS Functional Assessment Scale [18]: 33
- UHDRS Independence Scale [18]: 30
- HD Activities of Daily Living (ADL) Scale [32]: 2
- Emotional/Behavioral Functioning

**Observer Reported Outcomes (ObsRO)**
- Neuropsychiatric Inventory [128]: 2
- HD-ADL Scale [32]; Activities of Daily Living Scale [129]; Cognitive Behaviour Rating Scale [130]: 1

**Other Outcome Measures**
- Biomarkers (serum, plasma, CSF levels); Physiology; Neuroimaging

* = Five of these studies only used a single item (n = 1 study used the Luria Hand Position Item and n = 4 studies used the Maximal Chorea Item).

** = The study did not specify if the Barthel Index was administered as a ClinRO or a PRO.

UHDRS Motor Exam [18] was the most frequently used assessment of motor function in HD studies in general, as well as the clinical trials reviewed here (44% of ClinRO HD assessments used). Although this measure has some noted weaknesses, it has been used extensively in HD clinical trials, and has received a “core” recommendation as a motor functioning measure from the NINDS HD CDE working group [8]. We
<table>
<thead>
<tr>
<th>Outcomes measure</th>
<th>Brief description</th>
<th>NINDS rec</th>
<th>Psychometric data summary</th>
<th>Combined recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CLINICIAN REPORTED OUTCOMES (ClinRO)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Motor Functioning</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UHDRS Motor Exam [18]</td>
<td>15-item exam provides standardized ratings of ocularmotor function, dystonia, chorea, dystonia, gait, postural stability, and other parameters</td>
<td>CORE</td>
<td>• support for internal consistency (Chronbach's alpha = .95) [18] • support for responsiveness to change [48, 60, 131, 132] • items criticized for redundancy, scoring difficulty, and high cognitive loading [133]</td>
<td>RECOMMENDED</td>
</tr>
<tr>
<td>Abnormal Involuntary Movement Scale (AIMS) [19]</td>
<td>12-item scale designed to measure tardive dyskinesia</td>
<td>NOT REVIEWED</td>
<td>• support for reliability in elderly population [13] • moderate support for reliability and concurrent validity in clinical populations [135] • limited support for interrater reliability [136]; interrater reliability can be problematic in individuals that are not experienced at using this scale [137] • no reliability or validity data published for HD</td>
<td>NOT RECOMMENDED</td>
</tr>
<tr>
<td>Marsden &amp; Quinn Chorea Severity Scale [20]</td>
<td>developed based on unpublished observations, and modified based on the experiences of the authors</td>
<td>NOT REVIEWED</td>
<td>• no published data to support (or refute) its psychometric properties • almost exclusively utilized by the developers, with only one other published study using this measure [138]</td>
<td>NOT RECOMMENDED</td>
</tr>
<tr>
<td>Quantitative Neurologic Examination [32]</td>
<td>48-item neurological exam with two subscales: the Chorea Scale (involuntary movements), and the Motor Impairment Scale (MIS; voluntary movements)</td>
<td>NOT REVIEWED</td>
<td>• interrater reliability high (r = 0.95) and test-retest reliability good (r = 0.89) [21] • support for validity in HD [21, 139] • developers indicate that this scale may more accurately reflect HD severity, rather than voluntary movement [21]</td>
<td>NOT RECOMMENDED</td>
</tr>
<tr>
<td><strong>Cognitive Functioning</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UHDRS Cognitive Exam [18]</td>
<td>Verbal Fluency [22]</td>
<td>Verbal Fluency [22] requires participants to think of as many words that start with a particular letter; assesses executive function, semantic knowledge and word generation</td>
<td>SUPPLEMENTAL</td>
<td>• no support for responsiveness to change in HD samples [8]</td>
</tr>
</tbody>
</table>

Table 4: Summary of psychometric data and recommendations for measures used in HD pharmacological trials
<table>
<thead>
<tr>
<th>Outcomes measure</th>
<th>Brief description</th>
<th>NINDS rec</th>
<th>Psychometric data summary</th>
<th>Combined recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stroop [23]</td>
<td>Involves three trials: color naming, word reading and interference (naming color of word written in the wrong color ink); assesses executive functioning and response inhibition</td>
<td>CORE</td>
<td>• Support for responsiveness to change in prodromal and symptomatic HD [9, 40, 60, 140]</td>
<td>RECOMMENDED</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SUPPLEMENTAL</td>
<td>• Consistently reported as the most sensitive cognition measure in the HD literature (typically has the largest effect size of all cognitive measures), especially in prodromal HD [9, 40, 60, 131]</td>
<td>RECOMMENDED</td>
</tr>
<tr>
<td>Symbol Digit Modalities Test [24]</td>
<td>Requires pairing obscure symbols with numbers; provides an index of attention, visuospatial processing, working memory, and psychomotor speed</td>
<td>CORE</td>
<td>• Studies in other clinical populations support test-retest reliability [91, 141, 142], internal consistency [143], sensitivity [144], specificity [144], predictive validity [144, 146], and construct validity [145, 147, 148]</td>
<td>NOT REVIEWED</td>
</tr>
<tr>
<td>Mini-Mental State Exam [93]</td>
<td>30-item measure is designed to evaluate cognitive status</td>
<td>NOT REVIEWED</td>
<td>• Test bias related to age, education, and socioeconomic background [141, 149–155]</td>
<td>NOT RECOMMENDED</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Not sensitive to mild cognitive decline characteristic of prodromal HD [156–158]</td>
<td></td>
</tr>
<tr>
<td>Trailmaking Test [67]</td>
<td>Sequencing task: Trails A requires participant to sequentially connect numbers (assesses psychomotor speed); Trails B requires participant to sequentially alternate between numbers and letters (assesses psychomotor speed and executive functioning [159])</td>
<td>SUPPLEMENTAL</td>
<td>• In validation studies in HD</td>
<td>NOT RECOMMENDED</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Support in other clinical populations for reliability [160–162] and validity [22, 28, 163, 164]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Work in HD indicates it's sensitive in both prodromal and manifest HD [60]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Responsive to change supported for manifest HD but not prodromal HD [5,31]</td>
<td></td>
</tr>
<tr>
<td>Functional Limitations UHDRS Total Functional Capacity Scale [18]</td>
<td>5-item assessment that evaluates occupational, financial, domestic, self-care and level of care provided</td>
<td>CORE (dx only)</td>
<td>• Internal consistency supported [165] • Inter-rater reliability acceptable [165] • Support for validity [18, 166] • Support for responsiveness to change [88, 40, 31, 132, 167] • Support for sensitivity in prodromal HD [168]</td>
<td>RECOMMENDED</td>
</tr>
<tr>
<td>Outcomes measure</td>
<td>Brief description</td>
<td>NINDS rec</td>
<td>Psychometric data summary</td>
<td>Combined recommendation</td>
</tr>
<tr>
<td>------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------</td>
<td>-----------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>--------------------------</td>
</tr>
</tbody>
</table>
| UHDRS Functional Assessment Scale [18]               | 25 items designed to evaluate the participants’ ability to complete daily tasks   | CORE (dx only) | • support for internal consistency [18]  
• support for responsiveness to change in HD [33, 131]  
• support for sensitivity in prodromal HD [168]                                                                                                                  | RECOMMENDED              |
| UHDRS Independence Scale [18]                        | single item that reflects the participants’ level of independence                  | CORE (dx only) | • support for validity [18]  
• support for responsiveness to change [331, 132]  
• support for sensitivity in prodromal HD [168]                                                                                                                  | RECOMMENDED              |
| HD Activities of Daily Living                        | 17 items that evaluate observer reported adaptive functioning (there is both a     | NOT REVIEWED| • high internal consistency reported [32, 169]  
• support for concurrent and divergent validity [32, 169]  
• no responsiveness to change data                                                                                                                                    | NOT RECOMMENDED AT THIS TIME |
|                                                      | CliniRO and ObsRO version of this measure                                         |           |                                                                                                                                                                                                                                                                                        |                          |
| Emotional/Behavioral Functioning                     |                                                                                   |           |                                                                                                                                                                                                                                                                                        |                          |
| UHDRS Behavioral Exam [18]                           | 10 questions that evaluate frequency/ severity of mood, behavior, psychosis and    | RECOMMENDED| • support for internal consistency [18]  
• support for reliability and validity in HD samples [36, 38]                                                                                                                                         | NOT RECOMMENDED          |
|                                                      | obsession/obsessiveness                                                           |           |                                                                                                                                                                                                                                                                                        |                          |
| PBA-s (35)*                                          | 11 items based on the UHDRS Behavioral Exam; it has more specific guidance on the  | CORE      | • support for reliability and validity in HD samples [36, 38]  
• support for responsiveness to change in HD [39, 40]                                                                                                             | RECOMMENDED              |
|                                                      | administration and scoring                                                        |           |                                                                                                                                                                                                                                                                                        |                          |
| Hamilton Depression Rating Scale (Ham-D) [33, 34]    | 21-item measure designed to evaluate depression; it takes approximately 20–30      | NOT REVIEWED| • some support for internal consistency [30–79], interrater reliability [13, 177–184], and test-retest reliability [173, 179, 189]  
• support for concurrent and discriminant validity in other clinical populations [130–133, 180–187]  
• the total score is a weak index of depressive syndrome severity [198], and there are some deficiencies in content validity [173, 199]  
• low correlations with clinical assessments of depression [177, 200–204]  
• in HD, only some items able to discriminate between individuals with depressed mood from those without [41] | NOT RECOMMENDED          |
|                                                      | minutes to administer.                                                            |           |                                                                                                                                                                                                                                                                                        |                          |
| Clinical Global Impression Scale [19]                | 2-item measure designed to evaluate overall illness severity and global improvement | NOT REVIEWED| • some support for reliability [205] and validity [206, 307] in other psychiatric populations  
• no published data in HD to support its psychometric properties                                                                                                             | NOT RECOMMENDED AT THIS TIME |
<table>
<thead>
<tr>
<th>Outcomes measure</th>
<th>Brief description</th>
<th>NINDS rec.</th>
<th>Psychometric data summary</th>
<th>Combined recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PATIENT REPORTED OUTCOMES (PROs)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beck Depression Inventory (BDI-II) [42]</td>
<td>21-item self-report questionnaire designed to assess depressive symptomatology</td>
<td>NOT REVIEWED</td>
<td>• support for internal consistency [42, 170, 208–218], test-retest reliability [42], and concurrent and discriminant validity [42, 170, 190, 191, 211, 213, 216–225] in other clinical populations</td>
<td>NOT RECOMMENDED</td>
</tr>
<tr>
<td>Epworth Sleepiness Scale</td>
<td>8-items that assess excessive daytime sleepiness</td>
<td>NOT REVIEWED</td>
<td>• support for internal consistency [227, 228], test-retest reliability [227, 229–231], concurrent and discriminant validity [43, 228, 230–234] in other clinical populations</td>
<td>NOT RECOMMENDED</td>
</tr>
<tr>
<td>Hospital Anxiety and Depression Scale [44]</td>
<td>14-item measure designed to assess anxiety (n=7 items) and depression (n=7 items)</td>
<td>SUPPLEMENTAL</td>
<td>• psychometric support for both reliability [239–252], validity [240, 241, 243–245, 248, 251, 253, 254], and sensitivity/ specificity [170, 240, 244, 245] in other clinical populations</td>
<td>NOT RECOMMENDED AT THIS TIME</td>
</tr>
<tr>
<td><strong>OBSERVER REPORTED OUTCOMES (OBRs)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HD-ADL Scale [32]</td>
<td>17 items that evaluate observer reported adaptive functioning (there is both a ClinRO and ObsRO version of this measure)</td>
<td>NOT REVIEWED</td>
<td>• high internal consistency reported [32, 169]</td>
<td>NOT RECOMMENDED AT THIS TIME</td>
</tr>
</tbody>
</table>

* The PBA-s is the only measure that was not identified in the systematic review; it is included here because it is an improved version of the UHDRS Behavioral Exam (upon which it was based), i.e., the scoring and administration instructions were revised and clarified in this version, and because the NINDS CDE working group recommended this measure.
would also recommend this as a reasonable measure for inclusion in HD clinical trials.

Other commonly used ClinRO motor functioning measures are the Abnormal Voluntary Movement Scale (AIMS; 12% of the ClinRO assessments used) [19], the Marsden & Quinn Chorea Severity Scale [20] (used in 4% of the ClinRO assessments) and the Quantitative Neurologic Examination (QNE) [21] (used in three HD clinical trials). None of these three measures was recommended by the NINDS and psychometric support for these measures is lacking. We would not recommend their use in future HD clinical trials (see Table 4).

ClinRO cognitive functioning measures

Several ClinRO measures include cognitive assessments. As can be seen from Table 3, while many studies utilize cognitive measures, there are several more instances of sole use assessments in this category relative to any other type of measure. Below, we examine measures that have been used in 5% or more of the HD clinical trials. The three cognitive measures from the UHDRS that have been used with the greatest frequency in HD clinical trials are the Verbal Fluency Test (comprised 33% of the ClinRO assessments) [22], Stroop (comprised 31% of the ClinRO assessments used) [23], and Symbol Digit Modalities Test (comprised 27% of the ClinRO assessments used) [24]. While both the Symbol Digit and Stroop are recommended as core CDEs by the NINDS CDE groups, the verbal fluency test was given a supplemental classification, as it is not sensitive over time in HD samples [8]. We would recommend using the Stroop and the SDMT in HD clinical trials where cognition is the focus of treatment; we do not recommend using Verbal Fluency due both the lack of responsiveness to change data in HD, and the fact that there are other available cognitive measures that have support for responsiveness to change in HD samples.

In addition, the Mini Mental Status Exam (MMSE) was also administered as a measure of cognition for 15% of the ClinRO assessments, and the Trailmaking test [25] was used in 6% of the ClinRO assessments. Although there is validation data for the MMSE in other neurological populations [26] and the elderly [27], we were unable to identify published studies validating the MMSE in individuals with HD. The MMSE has been criticized poor discriminability among individuals with cognitive impairments [28, 30]. Thus, the MMSE is not a good candidate measure for inclusion in HD clinical trials. Furthermore, although the Trailmaking test has responsiveness to change for manifest HD, but not prodromal HD [5, 31], it is less desirable for inclusion in HD studies, especially those studies that wish to include the full spectrum of HD severity and/or symptomatology. We do not recommend this measure for use in HD.

ClinRO functional limitations measures

Several studies have also included measures of functional limitations from the UHDRS [18]: the Total Functional Capacity scale (TFC; included in 34% of the ClinRO assessments), the Independence Scale (used in 21% of ClinRO assessments, and the Functional Assessment Scale (included in 24% of ClinRO assessments). All three of these functional assessments were recommended by the NINDS HD CDE team as core measures in HD research [8], and have strong psychometric support. We recommend them as candidate measures for HD pharmacological trials.

In addition, the Huntington’s Disease Activities of Daily Living (HD-ADL) scale was used as a ClinRO assessment in two HD clinical trials [32]. While the developmental validation data suggests that this measure shows promise, additional data on the psychometric properties of this measure are likely needed before it is used consistently in HD research.

ClinRO Emotional/Behavioral functioning measures

ClinRO measures also include measures of emotional/behavioral functioning. Of note are the UHDRS behavioral exam [18] (used in 26% of the ClinRO assessments) and the HAM-D [33, 34] (used in 6% of the ClinRO assessments). While the UHDRS Behavioral Exam [18] includes recent psychometric data, the NINDS CDE group has recommended replacing this with the PBA-s [35] (which is based on the UHDRS behavioral exam, but has more detailed questions, more specific guidance on administration and scoring, and support for its reliability and validity in HD samples [36–38], as well as responsiveness to change over time [8, 39, 40]). We recommend the PBA-s in HD pharmacological trials moving forward.

The HAM-D [33, 34] has also been used in 6% of the HD ClinRO assessments. While there is some psychometric support for this measure in other clinical populations, there is also some concern. Furthermore, the one study that examined the HAM-D in HD found that although some items discriminated individuals...
with depressed mood from those without, several items could not; the authors therefore concluded that the HAM-D is of limited usefulness in HD [41]. We would not recommend using this measure in HD trials.

ClinRO “Other” assessments

Finally, 11% of HD ClinRO assessments included the Clinical Global Impression Scale [19], a measure designed to evaluate overall illness severity and global improvement (initially designed for use in schizophrenia research). As there is no published data in HD to support its psychometric properties, we do not recommend using this measure.

PRO measurements

The only PROs used in more than a single study were the Beck Depression Inventory Second Edition (BDI-II) [42], Epworth Sleepiness Scale [43] and the Hamilton Anxiety and Depression Scale (HADS) [44]. The Beck Depression Inventory, provides an assessment depressive symptomatology [42]. While there is data to support its psychometric properties in other clinical populations, the two studies have examined the psychometric properties of the BDI-II in HD were not very supportive. Taken together, there is not strong support for the utility of the BDI-II in individuals with HD. While there may be some support for the use of these measures in other clinical populations, the psychometric data for these measures in HD samples is lacking or limited. Furthermore, the NINDS does not make recommendations for any of these measures. We would not recommend these measures for inclusion in future HD clinical trials at this time.

As such, there is a need for PRO measures that have reliability, validity and responsiveness to change data to support their utility in HD pharmacological trials. This message is not new; previous work has highlighted the need for HD-specific PROs measures [8, 45]. For example, the NINDS CDE PRO working group highlighted a number of potentially useful measures that are either under development or recently developed, but have not yet received widespread use in HD. These include NIH-funded measurement development initiatives: PROMIS™ (www.nihpromis.org) [46, 47], Neuro-QOL (www.neuroqol.org) [48], and the HDQLIFE™ [49]. In addition, there are measures that have recently been developed in Europe that might also fill this void (i.e., the HD-Qol [50] and the HD Quality of Life Instrument [51]), although the HD-Qol has been criticized for not meeting statistical assumptions required for running item response theory [52], and the HD Quality of Life Instrument is only available in French and Italian [52]. Furthermore, since HD is a neurodegenerative condition, individual’s in the later stages often exhibit anosognosia, or a lack of insight into one’s own symptoms and deficits [53–55]. Anosognosia can compromise the reliability of a PRO, highlighting the importance of capturing information from another source (i.e., ClinRO and/or ObsRO). Thus, a more complete clinical symptom picture requires both types of information: a PRO and a complimentary ClinRO/ObsRO. Taken together there appears to be much work to be done in identifying a universally acceptable, psychometrically sensitive HD-specific PRO measure.

ObsRO measurements

ObsRO assessments were rarely included in HD clinical trials (only 3% of the HD clinical trials included and ObsRO measure) highlighting the need for additional work to identify existing measures, or develop new measures that may have utility in HD. One potential candidate for an ObsRo measure might include the HD-ADL Scale [32]; this measure has published ObsRo psychometric data in HD, and with additional work to confirm its psychometric properties, it might be appropriate for use in HD research.

Study limitations

This review has several limitations. First, to identify outcomes measures that were more likely to be responsive to an intervention, we focused our review on pharmacological/drug trials in HD. In this manner we did not review other experimental, interventional (e.g., non-pharmacological, device based) or observational studies. Therefore, future work is needed to fully evaluate the measures selected for inclusion in other intervention and observational studies, since some of these studies may use other measures that are sensitive to therapeutic effects in HD research. Furthermore, we only extensively reviewed measures that were used with some manner of frequency in HD clinical trials, and therefore, there are several other candidate measures that might have sensitivity in HD that were not evaluated. We also limited our search to clinical trials in English or Spanish and consequently may have missed instruments that are employed in other countries. Furthermore, much recent development work for HD PROs measures is underway, which would not have been captured.
CONCLUSIONS

This review provides a summary of the different outcomes measures being used in HD pharmacological trials. ClinRO measures are the most frequent outcomes measures in HD drug studies, and there are psychometric data to support the use of at least a few of these measures as COAs in HD clinical research. Contrary to this, much work still needs to be done before specific recommendations for HD PROs and ObsRO measures are made. To this end, there have been several efforts to develop HD PRO measures, although data are still needed across multiple research groups to support these new measures’ reliability, validity and sensitivity to change. Finally, ObsRO measures are lacking; additional development work in this area is needed if we are to utilize these types of measures in our HD pharmacological trials.

ACKNOWLEDGMENTS

This work was funded by the National Institute of Neurological Disorders and Stroke (NINDS) (Grant Numbers R03NS065194 and 1R01NS077946). The authors also wanted to express appreciation to Whitney Townsend, MLIS, who conducted the systematic literature searches.

CONFLICT OF INTEREST

The authors have no conflict of interest to report.

SUPPLEMENTARY MATERIAL

The supplementary tables are available in the electronic version of this article: http://dx.doi.org/10.3233/JHD-140115.

REFERENCES


N.E. Carlez... / Outcomes Measures in HD


