RESEARCH ARTICLE

Uses of the NIH Toolbox® in Clinical Samples

A Scoping Review

Rina S. Fox, PhD, MPH*, Manrui Zhang, MPH, MSW*, Saki Amagai, BA, Adrianna Bassard, MS, Elizabeth M. Dworak, MA, Y. Catherine Han, MS, Jessica Kassanits, BS, Corinne H. Miller, MLIS, Cindy J. Nowinski, MD, PhD, Amy K. Giella, BA, Jordan N. Stoeger, MEd, Kathleen Swantek, MLS, Julie N. Hook, PhD, MBA, and Richard C. Gershon, PhD

Correspondence
Dr. Gershon
gershon@northwestern.edu

Neurology: Clinical Practice August 2022 vol. 12 no. 4 307-319 doi:10.1212/CPJ.0000000000200060

Abstract

Background and Objectives

The NIH Toolbox® for the Assessment of Neurologic and Behavioral Function is a compilation of computerized measures designed to assess sensory, motor, emotional, and cognitive functioning of individuals across the life span. The NIH Toolbox was initially developed for use with the general population and was not originally validated in clinical populations. The objective of this scoping review was to assess the extent to which the NIH Toolbox has been used with clinical populations.

Methods

Guided by the Joanna Briggs Methods Manual for Scoping Reviews, records were identified through searches of PubMed MEDLINE, PsycINFO, ClinicalTrials.gov, EMBASE, and ProQuest Dissertations and Theses Global (2008–2020). Database searches yielded 5,693 unique titles of original research that used at least one NIH Toolbox assessment in a sample characterized by any clinical diagnosis. Two reviewers screened titles, abstracts, and full texts for inclusion in duplicate. Conflicts at each stage of the review process were resolved by a group discussion.

Results

Ultimately, 281 publication records were included in this scoping review ($n_{\text{Journal Articles}}$ = 104, $n_{\text{Conference Abstracts}}$ = 84, $n_{\text{Clinical Trial Registrations}}$ = 86, and $n_{\text{Theses/Dissertations}}$ = 7). The NIH Toolbox Cognition Battery was by far the most used of the 4 batteries in the measurement system ($n_{\text{Cognition}}$ = 225, n_{Emotion} = 49, n_{Motor} = 29, and $n_{\text{Sensation}}$ = 16). The most represented clinical category was neurologic disorders (n = 111), followed by psychological disorders (n = 39) and cancer (n = 31). Most (96.8%) of the journal articles and conference abstracts reporting the use of NIH Toolbox measures with clinical samples were published in 2015 or later. As of May 2021, these records had been cited a total of nearly 1,000 times.

Discussion

The NIH Toolbox measures have been widely used among individuals with various clinical conditions across the life span. Our results lay the groundwork to support the feasibility and utility of administering the NIH Toolbox measures in research conducted with clinical populations and further suggest that these measures may be of value for implementation in fast-paced clinical settings as part of routine practice.

*These authors contributed equally to this work.

Northwestern University (RSF, MZ, SA, AB, EMD, YCH, JK, CHM, CJN, AKG, JNS, KS, JNH, RCG), Chicago, IL; and University of Arizona (RSF), Tucson.

Funding information and disclosures are provided at the end of the article. Full disclosure form information provided by the authors is available with the full text of this article at Neurology.org/cp.

The NIH Toolbox® for the Assessment of Neurologic and Behavioral Function is a compilation of 47 brief, computerized, royalty-free measures designed to assess sensory, motor, emotional, and cognitive functioning.[1] The measurement system was initially commissioned by the NIH Blueprint for Neuroscience Research, a joint effort of 16 NIH Institutes.[2] It is appropriate for use across the life span of individuals, from ages 3 to 85 years, and is available in numerous languages. The NIH Toolbox is applicable for use in large-scale, longitudinal, epidemiologic studies because it enables the repeated assessment of these same domains of functioning across different developmental stages. In addition, it can provide a *common currency* to facilitate comparison of data across diverse study designs and populations.

When all measures included in the NIH Toolbox are administered as part of a comprehensive assessment battery, the measurement system can be completed in 2 hours or less.[3] This reflects the high priority the developers placed on lowering the burden of these measures to meet the needs of researchers designing large cohort studies. Thus, whenever possible, the NIH Toolbox measures were based on item response theory (IRT) and designed to be administered as computer adaptive tests (CATs). Other priorities for the NIH Toolbox development included the ability to measure multiple components of each of the 4 primary domains; versatile applicability across multiple study designs and populations; demonstration of psychometric and methodological strength across racial and ethnic groups, age ranges, and languages; sensitivity to developmental changes; and adaptability to emerging technologies.[1]

The NIH Toolbox was originally developed for research use with initial validation and norming limited to general population samples. It did not target specific disease outcomes nor was it intended to serve as an in-depth assessment of any domain of functioning or designed to be used as a diagnostic tool.[1] However, it is increasingly being used and evaluated in clinical populations. This is not surprising because many of the features that make the NIH Toolbox ideal for use in large-scale, longitudinal research also make it well-suited for use in clinical settings such as inpatient and outpatient medical clinics. For example, the NIH Toolbox measures' brevity and ease of use enhance their feasibility for integration into a busy clinical flow. Similarly, the ability to demonstrate sensitivity to change suggests these measures may be appropriate to track symptoms and functioning over the course of a disease or treatment trajectory and to discriminate between groups with and without clinically relevant functional deficits. However, the extent to which the NIH Toolbox has been used with clinical populations in research remains unknown. Thus, the objective of this scoping review was to assess uses of the NIH Toolbox in research with clinical populations as reported in the literature. By aggregating this information, the applicability of the NIH Toolbox measures to different clinical populations can be better understood and indexed.

Review Question

Our primary research question was as follows: In what clinical populations (i.e., populations defined by clinical categories) is the NIH Toolbox being used in research? We also explored which NIH Toolbox measures were used, general information about the study design and sample characteristics, funding sources, and publication effect of these records.

Methods

We conducted a scoping review[4] guided by the Joanna Briggs Methods Manual for Scoping Reviews.[5] Before embarking on the review, a preliminary search of MEDLINE, the Cochrane Database of Systematic Reviews, and JBI Evidence Synthesis was conducted, and no current or underway systematic reviews or scoping reviews on the topic were identified. The review methodology and results are reported in accordance with the PRISMA Extension for Scoping Reviews (PRISMA-ScR).[6]

Standard Protocol Approvals, Registrations, and Patient Consents

A protocol, publicly available at 10.18131/g3-2pvz-zb28, was prepared and indexed in advance.

Search Strategy

Articles were identified through database searches of PubMed MEDLINE, PsycINFO (EBSCO), ClinicalTrials.gov, EMBASE (Elsevier), and ProQuest Dissertations and Theses Global (January 2008 to June 2020). A search string was initially developed for PubMed MEDLINE (eTable 1, links.lww.com/CPJ/A359) and subsequently modified to enable searches within the remaining indexing systems. To capture as much of the available information as possible, conference proceedings, dissertations, and other gray literature were included, and there were no language restrictions.

Source of Evidence Screening and Selection

Following the search, all identified citations were collated and uploaded into the screening tool Rayyan.[7] Duplicate publications were removed before titles and abstracts were screened for inclusion. To pilot test and calibrate the search protocol, a subset of 70 titles was selected for initial screening by all reviewers (RSF, MZ, SA, AB, EMD, AKG, YCH, JK, CJN, JNS, KS, and JNH). These reviewers then met as a group to discuss the titles and resolve conflicts. Each of the remaining titles and abstracts was subsequently screened for inclusion by no fewer than 2 of these reviewers. Titles that were not excluded at this stage were retrieved in full and evaluated by no fewer than 2 reviewers against the inclusion criteria. All relevant full texts were accessible through institutional and/or public access platforms.

To be included in the review, studies were required to use at least one assessment from the NIH Toolbox in a sample characterized by any clinical diagnosis (i.e., identifiable by an ICD-10 code). No restrictions were imposed based on research setting (e.g., inpatient, outpatient, and community-based) or participant characteristics (e.g., sex and age). All types of original research were included, although meta-analyses, systematic reviews, editorials/opinions, and other review articles were excluded. Only studies published after 2008 were included in the review because that was the year that the first known article referencing the NIH Toolbox was published.[8]

Conflicts at each stage of the review process were resolved either by discussion between the 2 reviewers who screened the relevant title or, when conflicts could not be resolved through this approach, having a third reviewer screen the record to resolve the disagreement. For studies that did not include sufficient information to determine eligibility, attempts were made to contact the source authors. Following database study selection, the reference lists of included records were examined for additional studies that may be eligible for this review.

Whenever possible, multiple records reporting the results of the same study were linked, and data were extracted from only one of the records. For example, if an abstract was later published as a manuscript, the abstract was identified as a duplicate and excluded from the review results. This was only possible when records provided sufficient information to identify duplicates.

Data Extraction

Information was extracted from studies and recorded in a standardized data extraction form initially developed by RSF and iteratively refined based on feedback from all coauthors. In addition to the extraction plan, as outlined in the preregistered protocol, we also documented the total number of citations each record had received according to Google Scholar for journal articles and conference abstracts only. Sample characteristics were only extracted from journal articles because this information was not reliably included in other publication types.

The standardized data extraction form was initially piloted on 9 titles by the 8 reviewers who contributed to data extraction (RSF, MZ, SA, AB, EMD, YCH, JK, and KS). These reviewers then met as a group to discuss the titles and resolve conflicts, before data were extracted from each of the remaining full texts by no less than 2 of these 8 reviewers. The approach used to reconcile conflicts during screening was also used during extraction.

Following extraction, records were categorized into predefined categories based on the clinical diagnosis appropriate to describe study participants. These categories were originally selected to reflect those that were prioritized by the various NIH Institutes, Centers, and Offices that comprised the NIH Blueprint for Neuroscience Research when the NIH Toolbox was originally developed. Categories were refined by group consensus during the development of the standardized extraction form, as described earlier: (1) autoimmune diseases, (2) cancer, (3) cardiovascular diseases, (4) developmental disorders, (5) neurologic disorders, (6) obesity/diabetes/other metabolic syndromes, (7) psychological disorders, (8) rare diseases as defined by the National Organization for Rare Disorders,[9] (9) substance use, (10) transplant, and (11) other. Records were categorized as other if the clinical diagnosis characterizing the sample did not fit into any of the aforementioned 10 categories (e.g., osteoarthritis). Records presenting samples that met criteria for more than one category were included in all relevant categories. For example, a record presenting results from a sample of patients with Duchenne muscular dystrophy would be categorized as both a developmental disorder and a rare disease.[10]

Critical Appraisal

In accordance with the recommendations of the Joanna Briggs Institute, critical appraisal (i.e., assessment of risk of bias) was not conducted because the purpose of this scoping review was to identify and map the available evidence and not to provide a targeted answer to a clinical question.[5]

Data Availability

The bibliography of the 281 included publication records is summarized in eTable 2, links.lww.com/CPJ/A360.

Results

Search Results

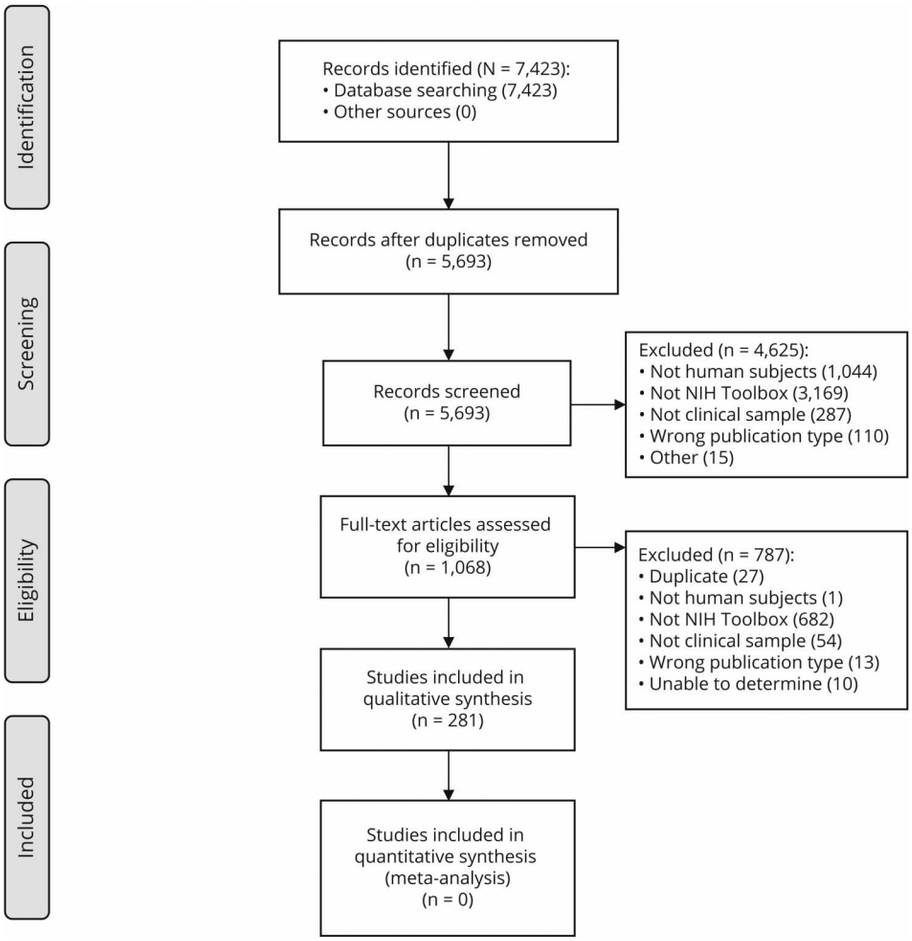
The search results are outlined in Figure 1. The database search yielded 7,423 titles and, after duplicates were removed, 5,693 titles and abstracts were screened for inclusion. Of these, 4,625 were excluded, while 1,068 full texts were retrieved and assessed for eligibility. Based on a full-text review, 787 studies were excluded, yielding a total of 281 records for inclusion in the review. No additional records were identified by searching the reference lists of the included articles.

NIH Toolbox Domains and Measures Used

Of the 281 records included in the review, 225 (80.1%) used the Cognition Battery, 49 (17.4%) used the Emotion Battery, 29 (10.3%) used the Motor Battery, and 16 (5.7%) used the Sensation Battery. Of note, 29 records administered measures from more than one domain (2 domains: n = 23, 3 domains: n = 2, 4 domains: n = 4). Within each of these categories, 203 (90.2%) records specified which measures of cognitive functioning were administered; 45 (91.8%) specified which measures of emotional functioning were administered; 21 (72.4%) specified which measures of motor functioning were administered; and 16 (100%) specified which measures of sensory functioning were administered.

More than half of the 225 records reporting the use of the Cognition Battery specified using measures of fluid cognitive

Figure 1 PRISMA Flowchart



abilities (Table 1). Specifically, studies reported using the Flanker Inhibitory Control and Attention Test (n = 181, 80.4%), Dimensional Change Card Sort Test (n = 166, 73.8%), List Sorting Working Memory Test (n = 162, 72.0%), Pattern Comparison Processing Speed Test (n = 160, 71.1%), and Picture Sequence Memory Test (n = 148, 65.8%). Measures of crystallized cognitive abilities including the Picture Vocabulary Test and the Oral Reading Recognition Test were used in 52.9% (n = 119) and 48.4% (n = 109) of records using the Cognition Battery, respectively. Few records reported use of the Oral Symbol Digit (n = 7, 3.1%) and Auditory Verbal Learning tests (n = 4, 1.8%); however, it is important to note that these 2 tests are considered supplemental and are not part of the core Cognition Battery.

Regarding the 49 records using Emotion Battery measures, slightly more used measures in the domains of Psychological Well-being and Social Relationships, including Emotional Support (n = 24, 49.0%), Positive Affect (n = 22, 44.9%), Life Satisfaction (n = 22, 44.9%), Friendship (n = 22, 44.9%), Loneliness (n = 20, 40.8%), and Perceived Rejection (n = 20, 40.8%), compared to measures of Negative Affect, including Sadness (n = 18, 36.7%), Fear (n = 18, 36.7%), and Anger (n = 17, 34.7%).

Finally, among those records reporting use of measures in the Motor and Sensation Batteries, the Standing Balance Test (n = 12, 41.3%) and the Pain Intensity Survey (n = 12, 75.0%) were the 2 most commonly used measures from the motor and sensation domains, respectively.

Clinical Diagnoses Represented

As summarized in Table 2, most of the records used an NIH Toolbox assessment among patients with neurologic disorders (n = 111, 39.5%), followed by patients with psychological disorders (n = 39, 13.9%), cancer (n = 31, 11.0%), cardiovascular diseases (n = 29, 10.3%), obesity/diabetes/other metabolic syndromes (n = 21, 7.5%), autoimmune diseases (n = 20, 7.1%), developmental disorders (n = 19, 6.8%), rare diseases (n = 11, 3.9%), substance use (n = 9, 3.2%), and transplant (n = 2, 0.7%). The other category (n = 40, 14.2%) contained a variety of clinical diagnoses,

Table 1 NIH Toolbox Measures by Publication Type

	Articles	Conference	Clinical trial registration	Dissertation	Total
Cognition (n = 203)					
Flanker	80	51	45	5	181
DCCS	75	43	41	7	166
LSWM	72	43	43	4	162
PCPS	74	42	39	5	160
PSM	67	40	37	4	148
Vocabulary	55	34	28	2	119
Reading	46	31	30	2	109
Oral symbol digit	5	0	2	0	7
Auditory verbal learning	4	0	0	0	4
Emotion (n = 45)					
Anger (affect, hostility, physical aggression)	7	3	7	0	17
Fear (affect, somatic, over anxious, separation anxiety)	7	3	7	1	18
Sadness	6	3	8	1	18
Positive affect	7	4	11	0	22
Meaning & purpose	7	3	9	0	19
Life satisfaction	7	4	11	0	22
Perceived stress	6	3	13	0	22
Self-efficacy	4	2	8	1	15
Emotional support	11	4	9	0	24
Instrumental support	9	4	6	0	19
Loneliness	11	4	5	0	20
Friendship	10	5	7	0	22
Perceived hostility	5	3	5	0	13
Perceived rejection	10	4	6	0	20
Apathy	0	0	2	0	2
Motor (n = 21)					
Standing balance	3	7	1	1	12
9-hole pegboard	5	2	1	1	9
Grip strength	3	3	1	1	8
2-minute walk endurance	2	2	1	0	5
4-meter walk gait speed	1	2	2	0	5
Sensation (n = 16)					
Pain intensity	3	3	5	1	12
Taste intensity	3	3	2	0	8
Odor identification	3	1	1	0	5
Pain interference	1	0	2	0	3

Continued

Table 1 NIH Toolbox Measures by Publication Type *(continued)*

	Articles	Conference	Clinical trial registration	Dissertation	Total
Words-in-noise	0	0	1	0	1
Visual acuity	0	0	1	0	1
Dynamic visual acuity	1	0	0	0	1
Vision-related quality of life	1	0	0	0	1

Abbreviations: DCCS = Dimensional Change Card Sort Test; LSWM = List Sorting Working Memory Test; PCPS = Pattern Comparison Processing Speed Test; PSM = Picture Sequence Memory Test; Vocabulary = Picture Vocabulary Test; Reading = Oral Reading Recognition Test; Flanker = Flanker Inhibitory Control and Attention Test.

including but not limited to osteoarthritis, fibromyalgia, insomnia, asthma, kidney disease, cystic fibrosis, vestibular dysfunction, hearing loss, liver cirrhosis, and survivors of critical illness. For most of the records included in our review (n = 233, 82.9%), the diagnosis characterizing study participants applied to only one category. This diagnosis was relevant to 2 or more categories in 48 records (2 categories: n = 45, 16.0%; 3 categories: n = 3, 1.1%).

While most records reported applications of the NIH Toolbox in research with clinical samples, 6 reported implementation in clinical settings.[11-16] Specifically, applications of NIH Toolbox Cognition Battery measures were reported among pediatric patients with brain tumors in both perioperative and follow-up inpatient and day treatment rehabilitation settings,[11-13] patients living with liver cirrhosis in an outpatient transplant clinic,[14] patients with dementia at a memory clinic,[15] and patients with celiac disease at a multidisciplinary outpatient gastroenterology clinic.[16]

Study Design

An observational study design was used in 39.9% (n = 112) of included records, compared with an experimental study design in 60.1% (n = 169). NIH Toolbox measures were primarily used as outcome variables (n = 250, 89.0%) and occasionally as independent (n = 9, 3.2%) or confounding (n = 6, 2.1%) variables. Of the 189 records that reported the country in which the research was conducted, most of the records (n = 169, 89.4%) indicated that research was completed in the United States. Other countries represented included Canada (n = 11, 5.8%), Australia (n = 2, 1.1%), the United Kingdom (n = 2, 1.1%), Zambia (n = 2, 1.1%), Israel (n = 1, 0.5%), Spain (n = 1, 0.5%), and Korea (n = 1, 0.5%). Only 78 records reported the language in which the NIH Toolbox assessment was administered. Of these, most of them administered NIH Toolbox measures in English (n = 70, 89.7%), while a few did so in Spanish (n = 7, 9.0%) and French (n = 1, 1.3%).

Regarding the number of times each NIH Toolbox assessment was used, 145 records (51.6%) reported administering NIH Toolbox measures only once, 75 records (26.7%) reported doing so twice, 33 records (11.7%) 3 times, and 16 records (5.7%) 4 or more times. Eight records reported adaptions in administration procedures.[17-24] Two adapted the test of taste intensity by using a different type of edible taste strips and by implementing tongue testing before whole mouth testing.[17,18] Six accommodated clinical populations with special needs by using nonstandard administration techniques for Cognition Battery measures, such as using a proctor, an alternate input device, or a nonstandard method for entering responses with the standard input device.[19-24] In addition, 9 records reported using modifications in scoring procedures.[25-33]

Sample Characteristics

Across all publication types, a notable proportion of records limited inclusion to either pediatric (i.e., upper age limit of 21 years, n = 60, 21.4%) or geriatric (i.e., lower age limit of 60 years, n = 48, 17.1%) samples. Of the 104 journal articles included in the review, sample sizes ranged from 2 to 1,002 (median = 50; interquartile range [IQR] = 25–157), and the mean age of study samples ranged from 4.7 to 75.1 years (median = 40.1; IQR = 28.0–54.6), as shown in Figure 2A and Figure 2B. The gender of participants who completed NIH Toolbox measures was specified in 91 journal articles: 2 journal articles included only male participants, whereas 13 articles included only female participants. As Figure 2C shows, the median sample size of female participants across all journal articles reporting this information was 29 (range = 0–594; IQR = 10–55), and the median sample size of male participants was 25 (range = 0–408; IQR = 7–61).

The number of participants from racial and ethnic minority groups (i.e., the number of non-White participants) was specified only in the 65 journal articles reporting results from studies conducted in the United States, Canada, and Great Britain. In these studies, the median sample size of racial and ethnic minorities was 24 (range = 0–282; IQR = 7–57), as shown in Figure 2D. The number of African American participants was specified in 58 journal articles. Across these, the median sample size of African American participants was 14 (range = 0–214; IQR = 4–31). The number of participants with Hispanic ethnicity was specified in 44 journal articles, and the median sample size of Hispanic participants was 7 (range = 0–232; IQR = 2–17).

Validation Studies

In 36 records, NIH Toolbox measures were validated against other legacy measures, many of which are typically considered

Table 2 Clinical Condition Categories by NIH Toolbox Domains

	All	Cognition	Emotion	Motor	Sensation
All publication types (n = 281)					
Neurologic disorders	111	93	17	13	3
Psychological disorders	39	34	11	0	0
Cancer	31	26	3	2	1
Cardiovascular diseases	29	24	8	6	2
Other	40	27	10	7	7
Obesity/diabetes/metabolic syndrome	21	17	2	2	2
Autoimmune diseases	20	12	1	2	2
Developmental disorders	19	17	2	0	0
Rare diseases	11	8	1	2	2
Substance use	9	5	2	0	1
Transplant	2	1	0	0	1
Journal articles (n = 104)					
Neurologic disorders	42	40	6	4	1
Psychological disorders	17	14	4	0	0
Cancer	15	13	1	1	1
Other	13	7	1	2	3
Cardiovascular diseases	11	10	4	2	1
Developmental disorders	7	7	1	0	0
Obesity/diabetes/metabolic syndrome	6	5	0	1	1
Autoimmune diseases	4	2	0	1	1
Rare diseases	4	3	1	0	1
Substance use	2	2	0	0	0
Transplant	1	0	0	0	1
Conference abstracts (n = 84)					
Neurologic disorder	37	28	3	7	0
Autoimmune diseases	11	6	1	4	0
Cardiovascular diseases	9	8	2	2	0
Cancer	8	6	1	1	0
Other	8	7	1	2	1
Obesity/diabetes/metabolic syndrome	7	6	1	1	0
Developmental disorders	7	6	0	0	0
Psychological disorders	5	5	1	0	0
Rare diseases	4	2	0	0	2
Substance use	2	2	0	0	0
Transplant	1	1	0	0	0
Clinical trial registration (n = 86)					
Neurologic disorders	31	24	8	2	2

Continued

Table 2 Clinical Condition Categories by NIH Toolbox Domains *(continued)*

	All	Cognition	Emotion	Motor	Sensation
Psychological disorders	17	15	6	0	0
Other	17	11	6	2	2
Obesity/diabetes/metabolic syndrome	8	6	1	0	1
Cancer	7	6	1	0	0
Cardiovascular diseases	7	4	3	0	0
Developmental disorders	5	4	1	0	0
Substance use	5	1	2	0	1
Autoimmune diseases	4	3	2	0	0
Rare diseases	3	3	0	0	0
Thesis/dissertation (n = 7)					
Cardiovascular diseases	2	2	0	0	0
Other	2	2	2	1	1
Autoimmune diseases	1	1	0	0	0
Cancer	1	1	0	0	0
Neurologic disorders	1	1	0	0	0

to be gold standard assessment tools. These included 33 measures of cognitive function, 2 measures of motor function, and one measure of vision. In these studies, patients with more than 18 clinical diagnoses were represented, including mild cognitive impairment, dementia, stroke, acquired brain injury, traumatic brain injury, Fragile X syndrome, Down syndrome, other intellectual disabilities, treatment-resistant psychosis, type II diabetes, and overweight/obesity. A complete list of the legacy measures against which NIH Toolbox assessments were validated, and the clinical diagnoses represented in these study samples, is summarized in eTable 3, links.lww.com/CPJ/A361.

Implementation of NIH Toolbox Measures in Clinical Practice

Six records, including 2 conference abstracts and 4 journal articles, have assessed the utility of implementing NIH Toolbox measures in research conducted in clinical settings as part of routine practice.[11-16] All 6 records reported using measures from the NIH Toolbox Cognition Battery. In particular, 3 records assessed the feasibility of administering the NIH Toolbox Cognition Battery to pediatric patients who had sustained acquired brain injuries as primary outcome measures.[11-13] These 3 records concluded that the NIH Toolbox Cognition Battery was a practical tool for tracking cognitive function among patients undergoing active treatment and those in rehabilitation because of the low patient burden and easy fit into the standard clinical workflow.[11-13] Of these, one record further reported successfully incorporating NIH Toolbox data into individualized clinical treatment plans.[12] The NIH Toolbox Cognition Battery has also been used as a diagnostic tool in a memory clinic to discriminate older adults with varying levels of cognitive impairment[15] and as a rapid cognitive screening tool in routine outpatient care to assess cognitive functioning among ambulatory patients with advanced liver disease.[14] Finally, the NIH Toolbox Cognition Battery has been used in a gastroenterology clinic to monitor treatment response in a case study that assessed the effect of a gluten-free diet on celiac disease.[16]

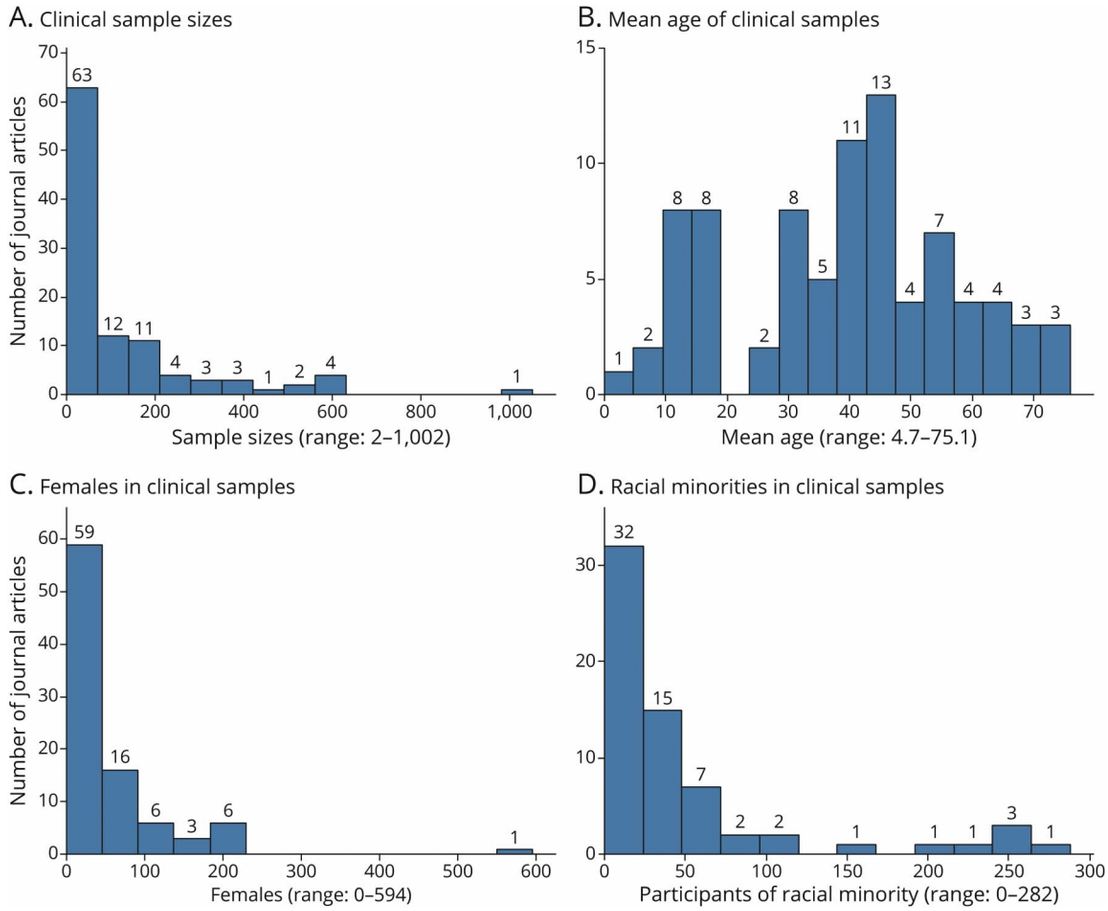
Publication Effect

Of the 281 records included in this scoping review, 104 were peer-reviewed journal articles (37.0%), 84 conference abstracts (29.9%), 86 registered records in clinicaltrials.gov (30.6%), and 7 theses/dissertations (2.5%). No records were identified that used NIH Toolbox measures with clinical samples between 2008 and 2010 (eFigure 1, links.lww.com/CPJ/A358). This number increased slowly from 2010 to 2015 and notably escalated after 2015. Similar trends were observed for the summed impact factors and Google Scholar citations for articles and conference abstracts by year (eFigure 1, links.lww.com/CPJ/A358). Across years, the median impact factor was 3.24 (Range = 0.6–23.6). Approximately one in 4 journal articles or conference abstracts (22.5%) was published in a journal with an impact factor ≥4. Of note, although most (96.8%) of the journal articles and conference abstracts using NIH Toolbox measures with clinical samples were published in 2015 or later, according to Google Scholar, as of May 2021 these records had already been cited a total of nearly 1,000 times.

Funding Source

Of the 120 publication records that reported a source of funding, 72.5% (n = 87) received funding from the NIH,

Figure 2 Sample Characteristics



27.5% (n = 33) from foundations, 24.2% (n = 29) from institutions, and 12.5% (n = 15) from non-NIH government agencies and 4.2% (n = 5) specifically reported having not received any funding. Most of these records reported only one funding source (n = 81, 67.5%), whereas some reported funding from multiple sources (2 sources: n = 30, 25.0%; 3 sources: n = 8, 6.7%; and 4 sources: n = 1, 0.8%).

Discussion

This is a study that overviews the use of NIH Toolbox measures in research with clinical populations. The findings from our scoping review indicate that there is a substantial and increasing body of literature demonstrating the uses of NIH Toolbox measures among patients across the life span with a wide range of clinical conditions. The results highlight the applicability of this measurement system across a variety of clinical populations and provide evidence to support their continued utility in future research.

The NIH Toolbox Cognition Battery was by far the most used of the 4 batteries in the measurement system, and neurologic conditions were the most represented clinical category included in this review. Although the NIH Toolbox was not developed to target specific disease outcomes, it is important to note that the relatively elevated use of the NIH Toolbox measures to assess cognition among patients with neurologic disorders does not contradict the central purpose of the measurement system. Rather, this may reflect the general applicability of cognitive testing to neurologic illness and highlights the value of the NIH Toolbox Cognition Battery as an assessment tool that can be applied in this clinical context. In addition, psychometric properties, such as construct validity, were evaluated in samples representing diverse clinical diagnoses, including patients with cognitive impairment, intellectual disabilities, brain injuries, and stroke. This strengthens the evidence supporting the applicability of the NIH Toolbox to a broad range of clinical populations.

Within the cognitive domain, 3 of the most commonly used measures were the Flanker Inhibitory Control and Attention Test, the Dimensional Change Card Sorting Test, and the Pattern Comparison Processing Speed Test. These are all tests of fluid cognition or one's ability to solve problems, think and act quickly, encode new memories, and perceive,

process, and respond to information in real time.[34] Fluid abilities are understood to be dynamic and are more susceptible to change than crystallized abilities, which reflect one's accumulated verbal knowledge and skill. Thus, these tests of fluid cognition may be particularly appropriate to detect deficits in cognitive function related to clinical conditions, which may be especially relevant to the practice of clinical neurology. Other tests included in the battery may be particularly valuable for other clinical populations and purposes. For example, the Standing Balance Test may have increased applicability to stroke survivors and the Odor Identification Test could be of special use among individuals with the SARS-CoV-2 virus. Therefore, it is important that researchers consider the specific needs of their individual research questions, study designs, and populations of interest when selecting NIH Toolbox measures for use in research.

Regarding study design, NIH Toolbox measures were administered not only in cross-sectional studies but also in longitudinal observational and experimental studies to measure changes over time. Several features of the NIH Toolbox make it a good fit for use in longitudinal research with clinical populations. The NIH Toolbox includes different forms of the same measure and uses IRT whenever possible to enable CAT.[35] Because items are dynamically selected each time based on the response to the previous item, practice and retest effects are greatly minimized.[36] Such practice and retest effects could threaten the internal validity of longitudinal studies where repeated administration of a cognitive test is needed. Relatedly, tests are able to achieve a high level of reliability with fewer items administered, reducing test burden. This is particularly relevant for populations for whom lengthy testing may be physically or mentally inappropriate and improves the feasibility of collecting data in a fast-paced clinical setting. This can help prevent attrition related to test burden in longitudinal research.

Our results showed that NIH Toolbox measures have been widely used with clinical populations representing sociodemographic diversity. Researchers evaluating pediatric and geriatric samples were primary users of NIH Toolbox measures, indicating the utility of NIH Toolbox in evaluating neuropsychological and neurophysiologic development across the life span. In addition, underrepresented and underserved populations such as patients with rare disorders and patients from racial and ethnic minority groups have also been fairly represented. Although most of the included records presented studies that were completed in the United States with English-speaking participants, additional records were identified from multiple countries reporting administration of NIH Toolbox measures in multiple languages. These findings are consistent with the original goal that the NIH Toolbox be applicable across diverse groups.[1,37] In the early stage of measure development, the NIH Toolbox project convened 5 working groups (i.e., pediatric, geriatric, cultural, non-English-speaking, and disabled) composed of internationally recognized experts in these fields.[37] These working groups provided recommendations regarding the instrument content and administration procedures, which were then incorporated into the development of the NIH Toolbox measures. Of note, while many of these recommendations have also benefited the implementation of the NIH Toolbox measures among clinical populations, specific adaptations for certain clinical populations may still be required. In our review, only 8 studies reported adjustments in administration procedures, mainly to enable populations with special needs to complete data entry (e.g., those with limited upper extremity mobility). However, it is important to note that this may be an underestimation. It is likely that some studies did not explicitly state when modifications were applied. While the NIH Toolbox provides an extensive list of potential adaptations,[19-24] future research might explore how those adaptations are implemented and interpreted in clinical settings. In addition, although this review did not identify any records reporting remote administration of NIH Toolbox measures or adaptations leveraging telemedicine approaches, such strategies are becoming more common and necessary, given the impact of the COVID-19 pandemic. Although not all NIH Toolbox measures can be administered remotely, many can. Specific guidance for remote administration of NIH Toolbox measures is available at nihtoolbox.org.

The included records highlight the scientific effect of the NIH Toolbox measurement system as used with clinical populations. Approximately two-thirds of included records were either peer-reviewed journal articles or conference abstracts. In total, these records have been cited nearly 1,000 times, and a quarter thereof were published in journals with an impact factor higher than 4. The first record documenting the use of an NIH Toolbox assessment with a clinical sample was released in 2010, 2 years after the first known publication introducing the NIH Toolbox[8] and the same year the measurement system was released for general use.[38] Thus, it is clear that the NIH Toolbox has been used with clinical populations since the measurement system was first introduced to the scientific community and this usage has rendered significant scientific influence over time.

Finally, although we sought to identify records using the NIH Toolbox in clinical research, we identified 6 records reporting applications of NIH Toolbox measures in clinical settings (e.g., inpatient and outpatient medical clinics) for clinical purposes (e.g., to inform clinical practice and/or as part of routine care). These records demonstrated that the NIH Toolbox is feasible for use at rehabilitation settings with a pediatric brain injury population,[11-13] is minimally disruptive to the clinical operations of a multidisciplinary outpatient gastroenterology clinic,[16] is valid for use among older adults at different stages of cognitive health at a memory clinic,[15] and can serve as a rapid screening tool in the outpatient setting to identify patients at a high risk for severe encephalopathy.[14] These findings underscore the great potential of the NIH Toolbox for use in routine clinical practice. Moreover, in addition to the use of IRT to enable CAT administration, which increases reliability while decreasing burden, an individual's score can be interpreted in the context of national norms calculated among typically developing

TAKE-HOME POINTS

- There is a substantial and increasing body of literature demonstrating the use of NIH Toolbox measures in sociodemographically diverse clinical populations.
- The NIH Toolbox Cognition Battery was the most used of the 4 batteries in the measurement system, and neurologic conditions were the most commonly represented clinical category.
- Within the cognitive domain, tests of fluid cognition were most used, including the Flanker Inhibitory Control and Attention Test, the Dimensional Change Card Sorting Test, and the Pattern Comparison Processing Speed Test.
- Evidence supported the utility and feasibility of administering NIH Toolbox measures in research conducted with clinical populations and suggested the potential value of using NIH Toolbox measures in fast-paced clinical settings as part of routine clinical practice.

individuals with similar sociodemographic characteristics.[39] Future research should continue exploring the implementation of NIH Toolbox measures in clinical settings and for clinical purposes.

This study has limitations. First, although we strictly followed the Joanna Briggs Methods Manual for Scoping Reviews, it is possible that we did not capture all the records that used the NIH Toolbox with clinical populations during our search time frame. Some conference proceedings and journal articles were excluded because they did not explicitly state that NIH Toolbox measures were used, although based on the description of the assessment tools, it was highly likely. When this occurred, we reached out to study authors as possible for confirmation, but nonetheless, we might have underestimated the number of studies that used NIH Toolbox measures among clinical populations. Second, this scoping review sought to describe the uptake of the NIH Toolbox in general and to highlight the extent to which the NIH Toolbox has been used with research in clinical samples. Because our results are intended to be descriptive, we did not attempt to evaluate any individual analysis or the methodological quality of included records nor did we summarize the conclusions of these studies. We suggest that clinicians and researchers review specific studies relevant to their individual scientific questions to capture this information. Moreover, future research could examine more nuanced uses of NIH Toolbox measures in specific clinical conditions. Third, although we attempted to link multiple records that seemed to present results of a single study, not all studies reported sufficient information to enable such linking. We presented our results by publication type to separate conference abstracts, journal articles, clinical trial registrations, and theses/dissertations associated with the same study. However, this cannot rule out the possibility that selected studies were included twice in separate records that presented different subsets of the same clinical sample. It is important to keep in mind that the results of this review have been summarized based on the number of publication records identified, not necessarily the number of studies presented.

NIH Toolbox measures have been widely used among individuals with various clinical conditions across the life span. Because the NIH Toolbox provides standardized measures as a common currency to be used across studies,[1] clinical researchers can meaningfully combine and compare evidence from different studies to systematically derive conclusions. Moreover, results lay the groundwork to support not just the applicability of the NIH Toolbox measures in clinical research conducted with clinical populations but also the potential utility of this measurement system for implementation in fast-moving clinical settings as part of routine practice.

Study Funding

This project was funded in part with Federal funds from the National Institute on Aging under grant number U2CAG057441 and The Environmental Influences on Child Health Outcomes (ECHO) program, Office of the Director, NIH, under award number U24OD023319. Dr. Rina S. Fox was supported by the National Cancer Institute under grant number K08CA247973.

Disclosure

The authors report no disclosures relevant to the manuscript. Full disclosure form information provided by the authors is available with the full text of this article at Neurology.org/cp.

Publication History

Received by *Neurology: Clinical Practice* November 15, 2021. Accepted in final form May 10, 2022.

Appendix Authors

Name	Location	Contribution
Rina S. Fox, PhD, MPH	Northwestern University, Chicago, IL; University of Arizona, Tucson, AZ	Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data; study concept or design; and analysis or interpretation of data
Manrui Zhang, MPH, MSW	Northwestern University, Chicago, IL	Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data; study concept or design; and analysis or interpretation of data

Continued

Appendix *(continued)*

Name	Location	Contribution
Saki Amagai, BA	Northwestern University, Chicago, IL	Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data
Adrianna Bassard, MS	Northwestern University, Chicago, IL	Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data
Elizabeth M. Dworak, MA	Northwestern University, Chicago, IL	Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data
Y Catherine Han, MS	Northwestern University, Chicago, IL	Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data
Jessica Kassanits, BS	Northwestern University, Chicago, IL	Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data
Corinne H Miller, MLIS	Northwestern University, Chicago, IL	Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data; study concept or design
Cindy J. Nowinski, MD, PhD	Northwestern University, Chicago, IL	Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data; study concept or design
Amy K. Giella, BA	Northwestern University, Chicago, IL	Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data
Jordan N. Stoeger, MEd	Northwestern University, Chicago, IL	Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data
Kathleen Swantek, MLS	Northwestern University, Chicago, IL	Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data
Julie N. Hook, PhD, MBA	Northwestern University, Chicago, IL	Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data; study concept or design
Richard C. Gershon, PhD	Northwestern University, Chicago, IL	Drafting/revision of the article for content, including medical writing for content; study concept or design; analysis or interpretation of data

References

1. Gershon RC, Wagster MV, Hendrie HC, Fox NA, Cook KF, Nowinski CJ. NIH toolbox for assessment of neurological and behavioral function. *Neurology.* 2013; 80(11 suppl 3):S2-S6.
2. NIH Blueprint for Neuroscience Research. Accessed July 12, 2022. neuroscienceblueprint.nih.gov.
3. The NIH Toolbox®. Accessed July 12, 2022. nihtoolbox.org.
4. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Method.* 2005;8(1):19-32.
5. Peters MDJ, Marnie C, Tricco AC, et al. Updated methodological guidance for the conduct of scoping reviews. *JBI Evid Synth.* 2020;18(10):2119-2126.
6. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med.* 2018;169(7):467-473.
7. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. *Syst Rev.* 2016;5(1):210.
8. Weintraub S, Tulsky DS, Dikmen S, Heaton R, Gershon R. P2-247: an introduction to the National Institutes of Health (NIH) toolbox for assessment of neurological and behavioral function: cognition domain. *Alzheimers Demen.* 2008;4(4S):T444.
9. NORD for Patients and Families. *Rare Disease Database*. Accessed February 11, 2021. rarediseases.org/for-patients-and-families/information-resources/rare-disease-information/.
10. Thangarajh M, Kaat AJ, Bibat G, et al. The NIH Toolbox for cognitive surveillance in Duchenne muscular dystrophy. *Ann Clin Transl Neurol.* 2019;6(9):1696-1706.
11. Watson W, Pedowitz A, Nowak S, Neumayer C, Kaplan E, Shah S. Feasibility of National Institutes of Health Toolbox Cognition Battery in pediatric brain injury rehabilitation settings. *Rehabil Psychol.* 2020;65(1):22-30.
12. Watson W, Selman J, Shah S, et al. A structured assessment protocol for children with acquired brain injuries in an inpatient rehabilitation hospital. *Arch Phys Med Rehabil.* 2017;98:e28.
13. Perkins S, Rubin J, Schlaggar B, et al. PDCT-13. Assessment of neurocognition and brain connectivity in pediatric brain tumor patients. *Neuro Oncol.* 2017;19(suppl 6):vi186.
14. Kim M, Liotta EM, Zee PC, et al. Impaired cognition predicts the risk of hospitalization and death in cirrhosis. *Ann Clin Transl Neurol.* 2019;6(11):2282-2290.
15. Hackett K, Krikorian R, Giovannetti T, et al. Utility of the NIH Toolbox for assessment of prodromal Alzheimer's disease and dementia. *Alzheimers Dement (Amst).* 2018;10:764-772.
16. Cassisi JE, Ross EJ, Vivier H, James N, Su LC. The impact of a gluten-free diet on celiac disease: a comprehensive evaluation of two cases using NIH patient reported outcome measures (PROMIS, NTCB, and Neuro-QoL). *J Clin Psychol Med Settings.* 2020;27:444-453.
17. Arlt JM, Smutzer GS, Chen EY. Taste assessment in normal weight and overweight individuals with co-occurring Binge Eating Disorder. *Appetite.* 2017;113: 239-245.
18. Coldwell S, Drangsholt M, Huggins K, et al. Reliability of a brief spatial test for assessment of gustatory function. *Chem Senses.* 2011;36:A24.
19. Dudley-Javoroski S, Lee J, Shields RK. Cognitive function, quality of life, and aging: Relationships in individuals with and without spinal cord injury. *Physiother Theory Pract.* 2020;38(1):36-45.
20. Lee J, Dudley-Javoroski S, Shields RK. Motor demands of cognitive testing may artificially reduce executive function scores in individuals with spinal cord injury. *J Spinal Cord Med.* 2021;44(2):253-261.
21. Lundine JP, Harnish SM, McCauley RJ, Zezinka AB, Blackett DS, Fox RA. Exploring summarization differences for two types of expository discourse in adolescents with traumatic brain injury. *Am J Speech Lang Pathol.* 2018;27(1): 247-257.
22. Magasi S, Marniss M, Gohil A. Reasonable accommodations affect the validity of scores on the NIH Toolbox Cognition Battery among people with neurological disabilities. *Qual Life Res.* 2015;24:9.
23. Magasi S, Harniss M, Tulsky DS, Cohen ML, Heaton RK, Heinemann AW. Test accommodations for individuals with neurological conditions completing the NIH Toolbox—Cognition Battery: an evaluation of frequency and appropriateness. *Rehabil Psychol.* 2017;62(4):455.
24. Shields RH, Kaat AJ, McKenzie FJ, et al. Validation of the NIH Toolbox Cognitive Battery in intellectual disability. *Neurology.* 2020;94(12):e1229-e1240.
25. Carlozzi NE, Goodnight S, Umlauf A, et al. Motor-free composites from the National Institutes of Health Toolbox Cognition Battery (NIHTB-CB) for people with disabilities. *Rehabil Psychol.* 2017;62(4):464-473.
26. Elias MN. The relationship between sleep quality and motor function in hospitalized older adult survivors of critical illness. 2018. [Doctoral dissertation, University of South Florida]. Digital Commons @ University of South Florida.
27. Farnsworth LK, Gilsanz P, Lacy ME, et al. Social support and cognitive function in type 1 diabetes: findings from the study of longevity in diabetes (SOLID) study. *Alzheimers Demen.* 2019;15(7):P1244.
28. Fee RJ, Montes J, Hinton VJ. Executive functioning in the dystrophinopathies and the relation to underlying mutation position. *J Int Neuropsychol Soc.* 2019;25(2): 146-155.
29. Hood AM, King AA, Fields ME, et al. Higher executive abilities following a blood transfusion in children and young adults with sickle cell disease. *Pediatr Blood Cancer.* 2019;66(10):e27899.
30. Lang S, Cadeaux M, Opoku-Darko M, et al. Assessment of cognitive, emotional, and motor domains in patients with diffuse gliomas using the National Institutes of Health Toolbox Battery. *World Neurosurg.* 2017;99:448-456.
31. Husein MM, McClintock S, Arslanagic E, Cullum M, Bernstein I Dewey R. The National Institute of Health (NIH) Toolbox: preliminary report of its use in Parkinson's disease. *Neurodegenerative Diseases.* 2011;8 (Suppl1)

32. St. Jude Children's Research Hospital. *Transcranial Direct Current Stimulation of The Temporal Cortex in Survivors of Childhood Acute Lymphoblastic Leukemia (ALL)*. 2019. Accessed February 19, 2021. ClinicalTrials.gov/show/NCT04105530.
33. St. Jude Children's Research Hospital. *Treatment for Executive Dysfunction in Adult Survivors of Childhood Acute Lymphoblastic Leukemia*. 2015. Accessed February 19, 2021. ClinicalTrials.gov/show/NCT02336282.
34. Heaton RK, Akshoomoff N, Tulsky D, et al. Reliability and validity of composite scores from the NIH Toolbox Cognition Battery in adults. *J Int Neuropsychol Soc*. 2014;20(6):588-598.
35. Weintraub S, Dikmen SS, Heaton RK, et al. Cognition assessment using the NIH Toolbox. *Neurology*. 2013;80(11 suppl 3):S54-S64.
36. Jones RN. Practice and retest effects in longitudinal studies of cognitive functioning. *Alzheimers Dement (Amst)*. 2015;1(1):101-102.
37. Victorson D, Manly J, Wallner-Allen K, et al. Using the NIH Toolbox in special populations: considerations for assessment of pediatric, geriatric, culturally diverse, non-English-speaking, and disabled individuals. *Neurology*. 2013;80(11 suppl 3):S13-S19.
38. Gershon RC, Cella D, Fox NA, Havlik RJ, Hendrie HC, Wagster MV. Assessment of neurological and behavioural function: the NIH Toolbox. *Lancet Neurol*. 2010;9(2):138-139.
39. Akshoomoff N, Newman E, Thompson WK, et al. The NIH Toolbox Cognition Battery: results from a large normative developmental sample (PING). *Neuropsychology*. 2014;28(1):1-10.