ANNALS of Clinical and Translational Neurology

Open Access

AMERICAN NEUROLOGICAL ASSOCIATION

RESEARCH ARTICLE

Reproducibility of cognitive endpoints in clinical trials: lessons from neurofibromatosis type 1

Jonathan M. Payne[1,2], Stephen J. C. Hearps[1], Karin S. Walsh[3], Iris Paltin[4], Belinda Barton[5,6,7], Nicole J. Ullrich[8], Kristina M. Haebich[1], David Coghill[1,2], Gerard A. Gioia[3], Alan Cantor[9], Gary Cutter[10], James H. Tonsgard[11], David Viskochil[12], Celiane Rey-Casserly[13], Elizabeth K. Schorry[14], Joseph D. Ackerson[15], Laura Klesse[16], Michael J. Fisher[4], David H. Gutmann[17], Tena Rosser[18], Roger J. Packer[3], Bruce Korf[19], Maria T. Acosta[3,20] & Kathryn N. North[1,2] for the NF Clinical Trials Consortium

[1]Murdoch Children's Research Institute, Royal Children's Hospital, Parkville, Victoria, Australia
[2]Department of Paediatrics, Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Melbourne, Victoria, Australia
[3]Center for Neuroscience and Behavioral Medicine, Children's National Health System, Washington, DC
[4]Division of Oncology, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania
[5]Kids Neuroscience Centre, The Children's Hospital at Westmead, Westmead, New South Wales, Australia
[6]Children's Hospital Education Research Institute, The Children's Hospital at Westmead, Westmead, New South Wales, Australia
[7]The University of Sydney Children's Hospital Westmead Clinical School, University of Sydney, Westmead, New South Wales, Australia
[8]Department of Neurology, Boston Children's Hospital, Boston, Massachusetts
[9]Department of Preventative Medicine, University of Alabama at Birmingham, Birmingham, Alabama
[10]School of Public Health, University of Alabama at Birmingham, Birmingham, Alabama
[11]Division of Neurology, The University of Chicago Medicine Comer Children's Hospital, Chicago, Illinois
[12]Department of Genetics, University of Utah, Salt Lake City, Utah
[13]Center for Neuropsychology, Boston Children's Hospital, Boston, Massachusetts
[14]Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio
[15]Department of Psychology, University of Alabama at Birmingham, Birmingham, Alabama
[16]Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, Texas
[17]Department of Neurology, Washington University School of Medicine, St Louis, Missouri
[18]Department of Neurology, Children's Hospital of Los Angeles, Los Angeles, California
[19]Department of Genetics, University of Alabama at Birmingham, Birmingham, Alabama
[20]National Institutes of Health, National Human Genome Research Institute, Bethesda, Maryland

Correspondence
Jonathan M. Payne, Murdoch Children's Research Institute, Royal Children's Hospital, Flemington Road Parkville, VIC 3052 Australia. Tel: +61 3 9936 6761; Fax: +61 3 8341 6212; E-mail: jonathan.payne@mcri.edu.au

Funding Information
This work was supported by the United States Army Medical Research and Materiel Command, Office of the Congressionally Directed Medical Research Programs, Department of Defense Neurofibromatosis Research Program, Grant Number W81XWH-05-1-0615.

Received: 3 July 2019; Revised: 31 October 2019; Accepted: 1 November 2019

***Annals of Clinical and Translational Neurology* 2019; 6(12): 2555–2565**

doi: 10.1002/acn3.50952

Abstract

Objective: Rapid developments in understanding the molecular mechanisms underlying cognitive deficits in neurodevelopmental disorders have increased expectations for targeted, mechanism-based treatments. However, translation from preclinical models to human clinical trials has proven challenging. Poor reproducibility of cognitive endpoints may provide one explanation for this finding. We examined the suitability of cognitive outcomes for clinical trials in children with neurofibromatosis type 1 (NF1) by examining test-retest reliability of the measures and the application of data reduction techniques to improve reproducibility. **Methods:** Data were analyzed from the STARS clinical trial (n = 146), a multi-center double-blind placebo-controlled phase II trial of lovastatin, conducted by the NF Clinical Trials Consortium. Intra-class correlation coefficients were generated between pre- and post-performances (16-week interval) on neuropsychological endpoints in the placebo group to determine test-retest reliabilities. Confirmatory factor analysis was used to reduce data into cognitive domains and account for measurement error. **Results:** Test-retest reliabilities were highly variable, with most endpoints demonstrating unacceptably low reproducibility. Data reduction confirmed four distinct neuropsychological domains: executive functioning/attention, visuospatial ability, memory, and behavior. Test-retest reliabilities of latent factors improved to acceptable levels for clinical trials. Applicability and utility of our model was demonstrated by homogeneous effect sizes in the reanalyzed efficacy data. **Interpretation**: These data demonstrate that single observed endpoints are not appropriate to determine efficacy, partly accounting for the poor test-retest reliability of cognitive outcomes in clinical trials in neurodevelopmental disorders. Recommendations to improve reproducibility are outlined to guide future trial design.

Introduction

Advances in genetically modified animal models and induced pluripotent stem cells have allowed detailed examination of the molecular pathways underlying cognitive deficits in neurodevelopmental conditions. This in turn has resulted in candidate molecular targets for therapeutic drugs, many of which have reversed behavioral phenotypes in animal models.[1–3] Despite a significant number of clinical trials targeting cognitive deficits in patients with genetic conditions, the vast majority have failed to translate the encouraging findings from preclinical trials.[4] The challenges of translating targeted treatments in human randomized controlled trials (RCTs) are typified by neurofibromatosis type 1 (NF1),[5–7] an autosomal dominant genetic condition associated with cognitive deficits affecting 80% of children.[8] Preclinical trials have identified several promising therapeutic targets. Lovastatin normalizes synaptic plasticity and rescues the learning phenotype in *Nf1*$^{+/-}$ mice by targeting RAS activation,[1] while methylphenidate and L-dopa improve attention by normalizing dopamine homeostasis in an *Nf1*$^{+/-}$ strain with bi-allelic inactivation in neuroglial progenitor cells.[9] Similarly, pharmacological blockade of Pak1 function within the amygdala rescues social deficits in *Nf1*$^{+/-}$ mice[10] and the HCN channel agonist lamotrigine rescues electrophysiological deficits by normalizing excitability of inhibitory interneurons.[11] Despite the optimism provided by these preclinical studies, attempts at translating findings to human clinical RCTs have been mixed (Table 1).[5–7,12–16] While benefits of methylphenidate for attention deficit/hyperactivity disorder (ADHD) have been reported,[16] and a very small study has suggested an effect of lovastatin on learning and memory,[15] three larger trials have failed to demonstrate any cognitive benefits of statin therapy.[5–7]

Translation from mouse studies to effective human clinical trials has proven difficult.[4,17] Potential explanations for this include inadequacy of animal models to recapitulate the complexity of the human disease state[18,19] and the frequent use of small sample sizes,[4] a common issue of RCTs in rare genetic conditions. It is also vital that cognitive endpoints are sensitive and reliable.[20] Here, we evaluated the suitability of various cognitive and behavioral endpoints for clinical trials in children with NF1 by analyzing data from the largest clinical trial for NF1-related cognitive deficits; the NF Clinical Trials Consortium's STAtin Randomized Study (STARS). This multicenter (11 sites), double-blind, randomized, placebo-controlled, parallel arm trial evaluated the efficacy of lovastatin for cognitive and behavioral deficits in children with NF1.[5] Similar to a previous study,[20] our initial aims were to (1) determine the severity of deficits on cognitive and behavioral outcomes at baseline and (2) examine the test-retest reliability of outcome measures after the 16-week treatment period. However, we extend this by also (3) investigating the utility of data reduction using confirmatory factor analysis (CFA) to improve psychometric properties of cognitive and behavioral outcomes, and (4) re-analyzing efficacy data using CFA-derived latent variables.

Methods

Participants

We analyzed data from the 144 children randomized in the STARS clinical trial.[5] Participants were treated with lovastatin (n = 74, mean age 11.5 years, 58% male) or placebo (n = 70, mean age 11.7 years, 62% male). Participants satisfied the clinical diagnostic criteria for NF1,[21] were aged between 8 and 15 years at screening, and demonstrated impaired performance on at least one primary outcome (≥1 SD below population mean). Exclusion criteria included full scale IQ < 70, symptomatic CNS pathology, significantly impaired vision/hearing, insufficient comprehension of English, low baseline total cholesterol (<90 mg/dL), and medications contraindicated on lovastatin. Children on a stable dose of stimulant medication (one month prior to screening and for study duration) were considered eligible.

The protocol was approved by the institutional review board at each site. The study was performed in accordance with the Declaration of Helsinki and Good Clinical Practice guidelines and was registered at clinicaltrials.gov (NCT00853580). Written informed consent was obtained from all parents/guardians before study entry.

Outcome measures

Efficacy outcomes were administered by psychologists at baseline and after 16-weeks of treatment and have previously been described in detail.[5] In brief, primary outcome measures were the Paired Associate Learning (PAL) test from the Cambridge Neuropsychological Automated

Table 1. Clinical trials targeting cognitive and behavioral outcomes in children with neurofibromatosis type 1.

Reference	Intervention (design)	*N* (Mean, SD)[2]	Key selection criteria	Treatment weeks	Primary outcomes	Secondary outcomes	Findings
5	Lovastatin (randomized, parallel group, placebo-controlled, double-blind)	146 (11.6, 2.1)	Impaired performance on primary outcome (s); FSIQ ≥ 70	16	CANTAB (PAL); TEA-Ch (Score)	CANTAB (SST, SWM, SOC); TEA-Ch (Sky Search, Creature Counting; Sky Search DT); CPT-II; COWAT; OA; BRIEF-P; JLO; BASC-2-P; BASC-2-SR; PedsQL-P; PedsQL-SR	Lovastatin did not improve visuospatial learning or attention
6	Simvastatin (randomized, parallel-group, placebo-controlled, double-blind)	62 (12.4, 10.4-14.4)[3]	Nil	12	RCFT-recall; Cancellation test; prism adaption; MRI ADC	RCFT-copy; Stroop CWT; BD and OA; Beery VMI; JLO	No significant effect of simvastatin except for improvement on OA
7	Simvastatin (randomized, parallel-group, placebo-controlled, double-blind)	84 (11.2, 7.9-16.0)[3]	No stimulant medication	52	CBCL-P; FSIQ	RCFT, Stroop CWI; Grooved Pegboard; CHQ; CBCL[4]	No significant improvement in cognitive deficits or behavioral problems on simvastatin
12	Methylphenidate (single arm, open-label)	20[1] (10.7, 2.2)	Diagnosed with comorbid ADHD	52	TOVA	CBCL-P; CBCL-T	Methylphenidate improved cognitive and behavioral attention; decline in aggressive behavior and social problems
13	Lovastatin (single arm, open-label, phase 1 escalation trial)	23 (13.0, 2.4)	History of learning disability; FSIQ > 80	12	TEA-Ch; CVLT; WRAML-2	JLO; Grooved Pegboard; BRIEF-P; BRIEF-SR; BASC-2-P; WASI	Some improvements in verbal and nonverbal memory after lovastatin use, but not in attention or vigilance
14	Simvastatin (randomized, parallel group, placebo-controlled, triple-blind)	30 (8.1, 1.8)	Diagnosed with comorbid ASD; *T* > 60 on SRS	12	ABC; CGI-S; Conners3; Resting state fMRI, MRI MRS (GABA, Glx), ADC, ASL, peripheral MAPK assay	Nil	Possible effect of simvastatin in brain regions associated with NF1 pathophysiology and social brain network
15	Lovastatin (randomized, parallel group, placebo-controlled, double-blind)	44 (25.7, 11.6)	FSIQ ≥ 70	14	BVMT-R; LNS	HVLT; Cancellation test; OA; CBCL-P; CBCL-SR; fMRI	Significant improvement on nonverbal declarative memory, working memory, and categorical verbal generation; significant improvement in internalizing symptoms for adult patients

(*Continued*)

Table 1. Continued.

Reference	Intervention (design)	*N* (Mean, SD)[2]	Key selection criteria	Treatment weeks	Primary outcomes	Secondary outcomes	Findings
16	Methylphenidate (randomized, crossover, placebo-controlled)	39 (9.3, 1.8)	FSIQ between 80-120; attention problems, school problems	9	SCRS-P	CRS-T; CDRS; CDI; STAI-C	Methylphenidate improved parent rated ADHD symptoms

Abbreviations: ABC, Aberrant Behavior Checklist; ADC, apparent diffusion coefficient; ADHD, attention deficit/hyperactivity disorder; ASD, autism spectrum disorder; ASL, arterial spin labelling; BASC-2-P, Behavior Assessment System for Children, Second Edition, Parent Form; BD, Wechsler Block Design; Beery-VMI, Beery Developmental Test of Visual Motor Integration; BRIEF-P; Behavior Rating Inventory of Executive Function, Parent Report; BRIEF-SR, Behavior Rating Inventory of Executive Function, Self-Report; BVMT, Brief Visuospatial Memory Test Revised; CANTAB, Cambridge Neuropsychological Test Automated Battery; CBCL-P, Child Behavioral Checklist, Parent Report; CBCL-SR, Child Behavioral Checklist, Self-Report; CBCL-T, Child Behavioral Checklist, Teacher Report; CDI, Children's Depression Inventory; CDRS, Children's Depression Rating Scale; CGI-S, Clinical Global Impression Scale; COWAT, Controlled Oral Word Association Test; CPT-II, Continuous Performance Test, Second Edition; CRS-T, Conners' Teacher Rating Scale; CVLT, California Verbal Learning Test; CWT, Color Word Interference test; fMRI, functional magnetic resonance imaging; FSIQ, Full Scale Intelligence Quotient; GABA, gamma-aminobutyric acid; Glx, glutamate + glutamine; JLO, Judgement of Line Orientation; LNS, Wechsler Letter Numer Sequencing; MAPK, Mitogen-activated protein kinase; NF1, neurofibromatosis type 1; OA, Wechsler Object Assembly test; PAL, Paired Associative Learning; PedsQL-P, Pediatric Quality of Life Inventory, Parent; PedsQL-SR, Pediatric Quality of Life Inventory, Self Report; RCFT, Rey Complex Figure Test; SCRS-P, simplified Conners' Parent Rating Scale; SOC, Stockings of Cambridge; SRS, Social Responsiveness Scale; SST, Stop Signal Task; STAI-C, State-Trait Anxiety Inventory for Children; SWM, Spatial Working Memory; TEA-Ch, Test of Everyday Attention for Children; TOVA, Test of Variables of Attention; WRAML-2, Wide Range Assessment of Memory and Learning, Second Edition.

[1] NF1 participants only.

[2] Years of age.

[3] Median (interquartile range).

[4] Teacher and youth self-report.

Battery (CANTAB) assessing visuospatial learning[22] and Score from the Test of Everyday Attention for Children (TEA-Ch) measuring sustained attention.[23] Secondary endpoints assessed attention, executive function and visuospatial skills, using Spatial Working Memory (SWM), Stockings of Cambridge (SOC), and the Stop Signal Task (SST) from the CANTAB, Sky Search, Sky Search DT and Creature Counting from the TEA-Ch, the Conners' Continuous Performance Task-II (CPT-II), Controlled Oral Word Association Test, Judgment of Line Orientation task, and Wechsler Object Assembly. In addition, parent behavioral rating scales assessed ADHD symptoms with the Conners-3, executive functions via the Behavior Rating Inventory of Executive Functions (BRIEF), internalizing behaviors on the Behavior Assessment System for Children-II (BASC-II), and psychosocial quality of life with the PedsQL. Participants also completed self-report versions of the BASC-II and PedsQL.

Data analysis

Unless stated otherwise, the complete case data set was used for all analyses. Baseline scores for all participants were used to establish the presence of pre-treatment cognitive and behavioral deficits. To allow direct comparisons between the various test scores, each standardized outcome was converted into a *z*-score such that 0 equals the reference group average with a standard deviation of 1. Differences between the NF1 group and normative reference data were tested using one-sample *t*-tests. To determine test-retest reliability of outcome measures, we examined associations between pre- and post-treatment scores of the placebo group using intraclass correlations.

We then conducted a rational reduction of cognitive and behavioral outcomes by creating a series of CFA models to determine the best-fitting model for the different cognitive domains (executive functioning/attention, memory, visuospatial) and behavior. CFA was chosen over exploratory factor analysis because it is a top-down, hypothesis-driven variant of structural equation modelling that is used to examine the expected connections between variables, treating each observed measure as one of multiple fallible manifest indicators of an underlying latent (i.e., unmeasured) construct. CFA models allow for random and systematic measurement error in the observed variables and the latent constructs can be interpreted with the measurement error taken into account. Scores were transformed prior to CFA such that a positive score reflected superior performance. The maximum likelihood method was used, and model assumptions examined (i.e., outliers, multivariate normality, linearity). Power calculations indicated approximately 90% power to achieve statistical significance in our proposed CFA model. Several models were tested and the best determined by the best overall fit indices including the χ^2 value, normed χ^2 (divided by the model degrees of freedom), Comparative Fit Index (CFI), and the root mean square error of approximation (RMSEA). Smaller χ^2 (and non-significant *P*-value), normed χ^2(<3) and RMSEA (<0.06) values and larger CFI (>0.9) values indicate a better fit. Standardized factor loadings of the best fitting model at baseline were used to calculate the scores for the latent factors at baseline and post-treatment.

As before, we determined test-retest reliability of latent factors by examining associations between pre- and post-treatment scores of the placebo group using intra-class correlations. We then re-examined the efficacy of lovastatin using the CFA-derived latent factors, comparing post-treatment scores of lovastatin and placebo conditions using analysis of covariance, adjusting for baseline scores. These intention to treat (ITT) analyses utilized data from participants with complete baseline data, applying multiple imputation for cases with missing 16 week data. Follow-up scores were imputed from baseline values, child age and sex, with 20 replications. Effect sizes were calculated (Cohen's *d*) to determine the magnitude of the group comparisons such that an effect size of 0.3 is small, 0.5 medium and 0.8 large.[24] All analyses were conducted in Stata IC (version 13.1).

Results

Summary baseline results for the intention to treat population are shown in Table 2. With the exception of child-reported internalizing behaviors ($P = 0.459$), baseline mean scores were significantly poorer than normative reference data (all, $P < 0.001$), confirming the presence of cognitive and behavioral deficits on these measures.

Test-retest reliability

Intra-class correlations between observed test values at baseline and post-treatment (week-16) for participants in the placebo condition are shown in Table 2. These test-retest effects allow examination of the stability of cognitive and behavioral outcomes of an inactive treatment within a realistic clinical trial setting. Higher correlations indicate higher stability and superior test-retest reliability of the outcome measure between pre-post assessments. While there are few standards for judging the minimum acceptable value for a test-retest reliability estimate,[25] reproducibility coefficients are typically considered adequate if greater than 0.7.[26] By these standards, test-retest reliability of the two primary outcomes was poor, with PAL and Score falling well below adequate levels (both, $r < 0.40$). Indeed, most secondary cognitive outcomes fell below adequate levels including widely used computerized tasks specifically designed to assay reliable change in clinical trials settings, including SWM and SST from the CANTAB. The SOC task in particular was very poor ($r = 0.10$). More traditional cognitive assessments tended toward higher pre-post correlations, but only the COWAT and JLO tasks demonstrated adequate reliabilities (both, $r > 0.72$). On the whole, behavioral questionnaire ratings tended towards higher pre-post correlations than cognitive outcomes. Parent-reported ADHD symptoms on the Conners-3 approached adequate levels, while reproducibility of parent-reported executive behaviors on the BRIEF was good. Parent and child-reported internalizing problems on the BASC-2 were adequate and psychosocial quality of life was just below acceptable levels for parent and child-report.

Confirmatory factor analysis

To reduce the number of variables and account for measurement error associated with observed variables, we performed a series of CFAs, based on theoretical models, using the cognitive factors of (1) executive functioning/attention, (2) memory, (3) visuospatial ability, and (4) general behavior on the complete-case population ($n = 104$) (Table S1). The best-fitting four-factor model is shown in Figure 1. Fit of this four-factor model was acceptable (χ^2 (154) = 181.59, $P = 0.06$; normed $\chi^2 = 1.18$; CFI = 0.95; RMSEA = 0.041, $P = 0.73$), model assumptions met, and factor loadings were significant (all, $P < 0.05$). Two cognitive variables (SOC mean number of moves and CPT commission errors) were excluded from the final model as their inclusion resulted in a poor fit (Table S2).

Results of pre-post intra-class correlations of the four latent cognitive and behavioral factors are shown in Table 3. As before, only the placebo condition was analyzed, employing ITT analysis with multiple imputation applied to those missing 16 week data. Results revealed noteworthy increases in test-retest reliability for the four latent variables compared to observed values reported in Table 2. The executive functioning/attention variable returned excellent reproducibility ($r = 0.83$), compared with the test–retest reliabilities of the eight observed cognitive values loading on the latent factor (range, 0.30-0.82). Pre-post correlations for the visuospatial and behavioral latent factors were also acceptable. The memory factor, consisting of PAL and SWM variables, fell slightly below the level of reliability considered adequate for clinical trials ($r = 0.66$).

Table 2. Total sample raw and standardized summary statistics for outcomes, and test–retest results within the placebo group.

	Total study population					Test-retest effects in placebo group only					
	Baseline						Baseline		16 weeks		
	N	Mean	SD	Z-score	SD	n	Mean	SD	Mean	SD	ICC
Cognitive test											
PAL[1]	143	16.1	19.4	−0.7	1.35	57	18.1	16.5	12.2	12.1	0.32
SWM[1]	139	47.1	16.2	−0.9	0.9	57	48.6	16.0	42.0	16.9	0.61
SWM strategy[1]	139	36.7	4.4	−0.7	1.0	57	37.0	4.7	35.5	5.5	0.51
SOC[1]	136	7.9	1.3	−0.5	1.1	53	8.1	1.3	7.9	1.4	0.10
SST[1]	135	252.2	89.8	N/A	N/A	53	235.7	79.5	228.1	94.7	0.51
Score[1]	136	5.9	2.2	−1.7	0.7	55	5.8	2.5	6.7	2.6	0.39
Sky search[1]	137	4.7	1.9	−0.7	1.0	55	5.2	2.2	4.6	2.2	0.63
Sky search DT[1]	136	10.0	17.2	−1.7	1.3	55	10.3	20.2	6.5	11.4	0.30
Creature counting[1]	137	3.9	2.2	−0.8	1.2	55	3.8	2.3	4.6	1.9	0.56
CPT omission[2]	135	59.7	15.2	−1.0	1.5	51	60.5	16.2	63.5	17.7	0.62
CPT commission[2]	135	55.6	9.1	−0.6	0.9	51	57.0	6.1	55.8	7.1	0.68
CPT RT[2]	135	55.0	14.2	−0.5	1.4	51	53.6	12.9	55.1	15.1	0.70
COWAT[1]	137	21.8	8.8	N/A	N/A	55	21.7	9.2	23.2	9.3	0.82
Object assembly[3]	137	6.8	3.1	−1.0	1.0	55	7.3	2.9	7.7	3.3	0.66
JLO[1]	135	14.9	6.2	N/A	N/A	54	14.5	5.7	16.4	6.3	0.73
Behavioral questionnaires											
ADHD inattentive[2]	137	64.7	13.4	−1.5	1.4	55	63.8	14.1	61.3	13.1	0.65
ADHD hyperactive/impulsive[2]	137	63.8	14.5	−1.4	1.6	55	63.5	15.3	62.9	16.9	0.68
BREIF BRI[2]	135	57.5	12.8	−0.8	1.3	55	56.4	13.5	54.9	12.6	0.79
BRIEF MCI[2]	135	63.5	11.4	−1.4	1.2	55	63.2	11.9	60.2	12.1	0.78
BRIEF GEC[2]	134	62.3	11.7	−1.2	1.2	55	61.6	12.1	58.9	12.5	0.80
Internalizing problems – parent[2]	133	54.4	12.3	−0.4	1.2	54	54.3	12.0	52.9	11.5	0.72
Internalizing problems – child[2]	132	50.8	9.4	−0.1	1.0	52	51.4	9.1	49.4	8.6	0.78
PedsQL psychosocial – parent[4]	133	64.0	17.0	−1.0	1.2	54	64.6	18.8	68.1	17.1	0.65
PedsQL psychosocial – child[4]	132	65.2	16.5	−1.0	1.1	54	62.7	17.1	67.2	17.7	0.69

Abbreviations: ADHD, attention-deficit/hyperactivity disorder; BRI, Behavioral Regulation Index; BRIEF, Behavior Rating Inventory of Executive Function; COWAT, Controlled Oral Word Association Test; CPT, Continuous Performance Test, Second Edition; DT, Divided Attention; GEC, Global Executive Composite; ICC, Intraclass correlation coefficient; M, Mean; MCI, Metacognition Index; PAL, Paired Associated Learning; RT, Reaction Time; SOC, Stockings of Cambridge; SST, Stop Signal Task; SWM, Spatial Working Memory; N/A, Normative reference data not available.

[1] Raw score.

[2] T score.

[3] Age scaled score.

[4] Summary score.

Having established superior psychometric properties of the CFA-derived latent variables, efficacy data from the STARS clinical trial were reanalyzed using the latent variables (Table 4). Again, this ITT analysis incorporated multiple imputed 16 week data where missing. While the analysis revealed no significant effect of lovastatin on any of the latent variables (all, $P > 0.17$), the consistency of effect sizes for the latent variables was considerably increased compared to those reported for the observed test scores. For example, lovastatin demonstrated a similar positive but non-significant effect on all three CFA-derived cognitive variables (d range; 0.24–0.26). In comparison, the observed test scores revealed noticeably more variation (d range; −0.23 to 0.36).[5]

Discussion

The selection of reliable outcomes that are sensitive to change is critical to the interpretation of any clinical trial, and is a particularly important issue for cognitive studies. Unreliable measures adversely affect the ability to detect positive responses to treatments and may explain the translational failure of mechanism-based clinical trials across a wide number of genetic conditions.[4] Here we present a detailed and methodical reexamination of the outcome data from STARS to advance the design of future cognitive clinical trials.

First, consistent with previously published data,[8,20] we demonstrated broad-ranging cognitive and behavioral

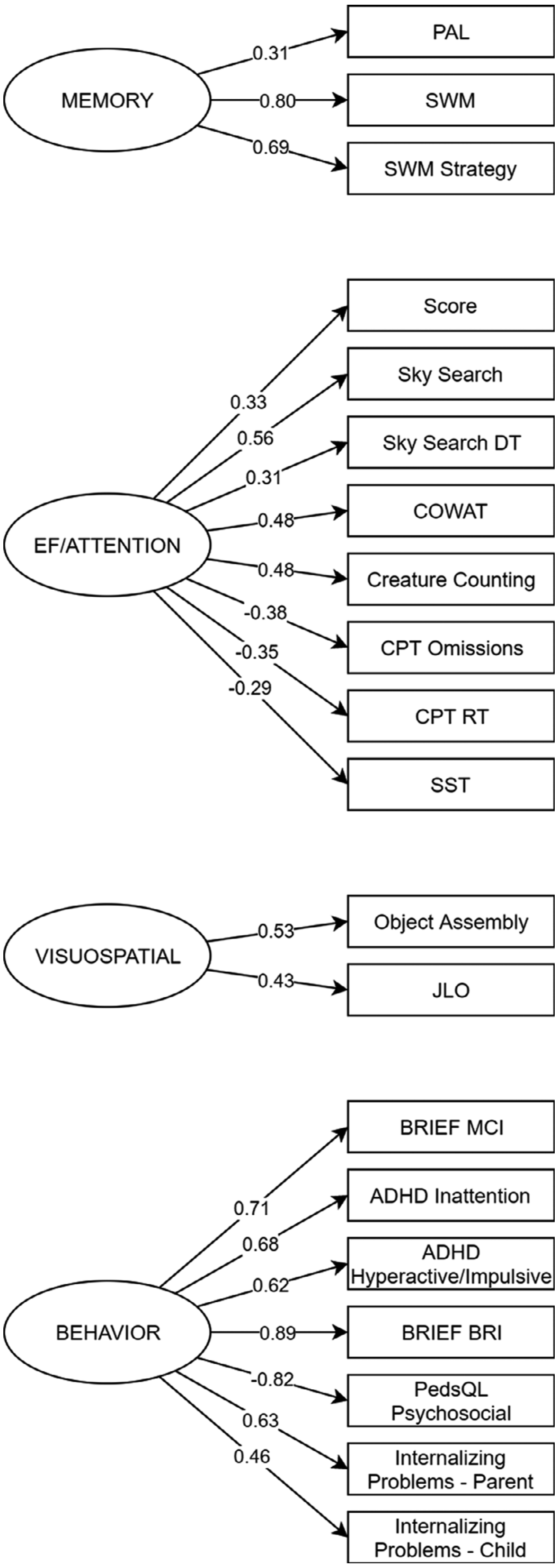

Figure 1. Data reduction for neuropsychological measures conducted using confirmatory factor analysis. The conceptual model shows how we hypothesized our observed variables (parallelogram) related to the four latent cognitive constructs (ellipse). For ease of presentation the diagram shows standardized factor loadings but does not contain error terms, cross loadings, or correlations between latent factors. ADHD, attention-deficit/hyperactivity disorder; BRI, Behavioral Regulation Index; BRIEF, Behavior Rating Inventory of Executive Function; COWAT, Controlled Oral Word Association Test; CPT, Continuous Performance Test, Second Edition; DT, Divided Attention; GEC, Global Executive Composite; ICC, Intraclass correlation coefficient; M, Mean; MCI, Metacognition Index; PAL, Paired Associated Learning; RT, Reaction Time; SOC, Stockings of Cambridge; SST, Stop Signal Task; SWM, Spatial Working Memory.

deficits in participants with NF1. All baseline cognitive outcomes were significantly below population reference data with particularly large deficits seen on sustained and divided attention measures from the TEA-Ch, confirming attention as a key cognitive deficit in NF1.[27] At the behavioral level, inattentive and hyperactive ADHD symptoms and executive behaviors were rated as the most significant problem areas. Consistent with previous evidence,[20] internalizing problems were less affected in our cohort.

Second, this study revealed poor reliability of commonly used cognitive outcomes in a real-life clinical trial setting. This was particularly striking for the CANTAB tasks, which all demonstrated test–retest reliabilities below acceptable levels, and in some cases, reproducibility was very poor. The variability noted across most cognitive tests was sufficiently substantial to compromise outcome stability, limiting the sensitivity of observed values to detect true change in this population. Importantly, reliabilities were considerably lower than published reliabilities in normative samples. As such, reliabilities of TEA-Ch subtests were markedly lower (range, 0.30–0.63) than those reported in a normative sample (range, 0.64–0.90).[23] While direct comparison of individual CANTAB tasks is more difficult, data from adult populations suggest higher reliabilities in neurotypical samples (range, 0.60–0.68) compared with our clinical trial (range, 0.10–0.61).[28]

These results raise important questions about the source of the poor reproducibility. One possible account is that executive tests of frontal lobe functioning are particularly sensitive to practice effects and are only valid when they are novel, since performance on these evaluations can rapidly improve as soon as an optimal strategy is discovered (practice effect), but will improve less if no strategy is found.[28,29] However, there is no clear explanation as to why strategy-driven practice effects would be differentially affected in children with NF1. Poor reliability was also observed on non-executive tests, including

Table 3. Test–retest effects of latent cognitive and behavioral domains from the intention to treat placebo group[1].

	Total complete case population			Test-retest effects, placebo group only				
	Baseline			Baseline		16 weeks		
	M	SD	*n*	*M*	SD	*M*	SD	ICC
Latent variable								
EF/attention	0.01	0.23	50	−0.02	0.25	−0.04	0.24	0.83
Visuospatial	0.04	0.38	50	−0.02	0.37	−0.08	0.41	0.78
Memory	0.02	0.48	50	−0.03	0.45	−0.07	0.39	0.66
Behavior	0.03	0.51	50	0.04	0.53	−0.02	0.53	0.84

Abbreviations: EF, Executive functioning; ICC, Intraclass correlation coefficient; M, Mean; SD, Standard deviation.

[1] Multiple imputation applied to all placebo cases with baseline data (20 replications).

Table 4. Between group comparisons at 16 week using latent cognitive and behavioral domains from the intention to treat population.

	Control					Lovastatin									
		Week 0		Week 16			Week 0		Week 16						
Outcome	*n*	*M*	SD	*M*	SD	*n*	*M*	SD	*M*	SD	*β*	95%CI	*P*	*d*	95%CI
Latent variable															
EF/attention	50	−0.02	0.25	−0.03	0.24	54	0.01	0.19	0.03	0.21	0.04	−0.02 to 0.09	0.174	0.26	−0.13 to 0.65
Visuospatial	50	−0.02	0.37	−0.08	0.40	54	0.04	0.39	0.04	0.41	0.07	−0.04 to 0.17	0.208	0.24	−0.14 to 0.63
Memory	50	−0.03	0.45	−0.07	0.40	54	0.03	0.50	0.07	0.55	0.11	−0.05 to 0.27	0.184	0.26	−0.12 to 0.65
Behavior	50	0.04	0.53	−0.03	0.54	54	0.00	0.47	0.00	0.53	0.05	−0.09 to 0.19	0.474	0.14	−0.24 to 0.53

Abbreviations: EF, Executive functioning; M, Mean; SD, Standard deviation; *d*, Cohen's *d*.

our two primary outcomes, arguing against a strategy effect. A further explanation for poor reproducibility in this study is that the characteristic attention deficits in NF1 result in variable performances and low stability across all cognitive measures.[27] In this instance, high variation in attention, impulse control, and perhaps motivation in clinical populations such as NF1 may critically reduce the reliability and validity of observed cognitive test scores such that they are unable to reliably quantify the cognitive domain the test is purported to assess.

Third, we found that a theory-based reduction of our cognitive and behavioral outcomes using CFA considerably improved the reproducibility of our endpoints. CFA in hypothesis-driven cognitive trials offers several advantages. In measurement theory, each observed cognitive test outcome reflects the underlying "true" score of the unmeasured latent construct with measurement error. CFA allows the loading of multiple observed values onto an underlying latent construct while modeling measurement error. As such, inferences about latent constructs can be interpreted as if measurement error was removed.[30] Once error was accounted for in our model, reproducibility of the executive functioning/attention, visuospatial, and behavioral latent factors exceeded acceptable levels for clinical trials. Critically, reproducibility of these latent factors improved despite the background of unreliable observed data.

Fourth, applicability and utility of our model was demonstrated by the consistent range of effect sizes across the reanalyzed efficacy data. Analysis of the three cognitive latent variables suggested small, uniform, but nonsignificant effects of lovastatin across all cognitive domains and that there is no clinical benefit of lovastatin. However, the consistency of treatment effects was considerably increased compared to previously reported effects using observed cognitive test scores,[5] demonstrating the advantages of data reduction and error modelling with CFA. The uniformity of treatment effects favoring lovastatin raise the possibility that statin therapy produced cognitive effects somewhat analogous to the results reported in preclinical trials,[1] but that the effect size was too small to yield statistical and clinical significance. While we do not recommend lovastatin as a treatment for cognitive deficits based on these results, our analyses suggest that proof-of-concept trials in genetic conditions such as NF1 should not be considered futile, and that variables, such as outcome measures utilized, sample size, as well as the dose, length and age of treatment, need to be carefully considered in pre-trial therapeutic analyses. Other novel surrogate outcome measures, such as biomarker assessments and functional neuroimaging,[31] may prove more sensitive to change than cognitive tests. Interestingly, the consistent effects across all cognitive domains also suggest that any benefits of targeted treatments in NF1 may not necessarily be confined to cognitive tasks attempting to replicate murine behavioral paradigms, or indeed to a particular cognitive domain. Rather, cognitive domains and outcomes selected should focus on areas of high clinical relevance to the NF1 population.

There are several critical lessons from this study which should be instructive to guide future cognitive clinical trials. The inherent variability on cognitive testing in clinical cohorts such as NF1 suggest that a single observed endpoint to estimate a cognitive ability is likely not appropriate. Poor reliability reduces statistical power, which calls for a substantial increase in sample size (not cost effective or feasible for rarer disorders), or an unacceptably high increase in the risk of false-positive or negative results, which may have resulted in false-positive findings in previous trials.[5,6,15] Before developing a clinical trial protocol, we strongly recommend piloting potential cognitive outcomes at multiple time points in an untreated clinical group to provide (1) base rates of deficits, (2) reproducibility of the specific cognitive endpoints and (3) valuable data on reliability to inform sample size calculations for clinical trials. Given specific cognitive tests are unlikely to return adequate reliability using observed test scores, investigators should consider applying targeted batteries of tasks and data reduction techniques, such as CFA, to decrease the number of comparisons made and reduce measurement error, thereby maximizing the reproducibility of cognitive outcomes.

It is also important to reflect on these recommendations within the context of the US Food and Drug Administration (FDA) Regulatory Standards used for approving new drugs. While specific guidance regarding observed versus latent performance outcome variables are not provided, the FDA stipulates that the study protocol and report of results should explain the variables measured, the methods of observation, and the criteria used to assess response. Importantly, the assessment methods should also be reliable.[32] Thus in order to retain the integrity of a priori reporting of methodology with the study protocol and on clinical trial registries, it is essential that the planned treatment of cognitive data be documented before trial commencement and that the choice of cognitive tasks, as well as their theoretical grouping, is guided by current conceptual models. It is also important that cognitive endpoints are sensitive to intervention. To this end, multifactorial scores that fully or partially rely on crystalized abilities, such as full scale IQ, are not recommended as cognitive endpoints for shorter-term trials (e.g., 16 weeks),[33] as the abilities they measure (learned knowledge and facts based on experience) are likely to require >12 months to demonstrate change.

When interpreting the current results, it is important to consider that test-retest reliability was assessed within a placebo condition. While it is possible that a genuine placebo effect influenced reliability coefficients,[34] our analyses retain the distinct advantages of examining data from a realistic clinical trial setting using a blinded, randomized sample. Importantly however, we have demonstrated here that despite any potential placebo effects, test-retest reliability of cognitive outcomes can be considerably improved through the application of careful, theory-driven data reduction methods.

In conclusion, this study highlights the importance of reliable cognitive endpoints in clinical trials. To improve translation from preclinical studies, future clinical trials will need to adopt methodologies that overcome the limitations of cognitive assessment. Although the results for this study are specific for NF1, the methods and recommendations discussed are likely to be relevant for other neurodevelopmental disorders, including autism and ADHD.

Acknowledgments

This work was supported by the United States Army Medical Research and Materiel Command, Office of the Congressionally Directed Medical Research Programs, Department of Defense Neurofibromatosis Research Program, Grant Number W81XWH-05-1-0615.

Conflict of Interest

JMP reports grants from the US Department of Defense and is supported by a Murdoch Children's Research Institute Clinician-Scientist Fellowship. IP reports grants from the US Department of Defense. DC reports grants and personal fees from Shire; personal fees from Eli Lilly, Medice, Novatis, Oxford University Press, Servier; grants from Vifor. GAG is a coauthor of the Behavior Rating Inventory of Executive Function (BRIEF) measure, published by Psychological Assessment Resources, Inc. (PAR), for which he receives royalties. GC reports personal fees (Data and Safety Monitoring Boards) from AMO Pharmaceuticals, Biolinerx, Horizon Pharmaceuticals, Merck, Merck/Pfizer, Opko Biologics, Neurim, Ophazyme, Sanofi-Aventis, Reata Pharmaceuticals, Receptos/Celgene, Teva pharmaceuticals, NHLBI (Protocol Review Committee), NICHD (OPRU oversight committee). He also reports personal fees (Consulting or Advisory Boards) from Atara Biotherapeutics, Axon, Biogen, Biotherapeutics, Argenix, Brainstorm Cell Therapeutics, Charleston Labs Inc, Click Therapeutics, Genzyme, Genentech, GW

Pharma, Klein-Buendel Incorporated, Medimmune, Medday, Novartis, Roche, Scifluor, Somahlution, Teva pharmaceuticals, TG Therapeutics, UT Houston. GC is also President of Pythagoras Inc., a private consulting company located in Birmingham AL. CR-C reports grants from the Children's Tumor Foundation during the conduct of the study. MJF reports grants from the US Department of Defense. DHG reports a patent Neurofibromatosis Gene with royalties paid, a patent Neurofibromin pathway modulators issued, and a patent Neurofibromin/dopamine signaling as a biomarker for cognitive and behavioral problems in children with neurofibromatosis type 1 pending.

Author Contributions

JMP contributed to study designed grant writing; helped develop study material including the protocol, played a major role in data acquisition; participated in data analysis and interpretation of results; and drafted and revised the manuscript. SJCH completed the statistical analysis; interpreted the data; and critically revised the manuscript for important intellectual content; KSW played a major role in acquisition of data; interpretation of results; and critically revised the manuscript for important intellectual content. IP interpreted the data; drafted the manuscript for intellectual content. BB contributed to study designed grant writing; interpreted the data; helped develop study material including the protocol, and critically revised the manuscript for important intellectual content. NJU contributed to study design; played a major role in acquisition of data; and critically revised the manuscript for important intellectual content. KMH, DC, AC and GC interpreted the data; and critically revised the manuscript for important intellectual content. GAG contributed to study designed grant writing interpreted the data; and critically revised the manuscript for important intellectual content. JTH, DV, CRC, EKS, JDA, LK, MJF, DHG, TR, RJP and BK played a major role in data acquisition; and critically revised the manuscript for important intellectual content. MTA contributed to study designed grant writing; helped develop study material including the protocol, played a major role in acquisition of data; interpreted the data; and critically revised the manuscript for important intellectual content. KNN contributed to study designed grant writing; helped develop study material including the protocol, played a major role in data acquisition; participated in data analysis and interpretation of results; and drafted and revised the manuscript.

References

1. Li W, Cui Y, Kushner SA, et al. The HMG-CoA reductase inhibitor lovastatin reverses the learning and attention deficits in a mouse model of neurofibromatosis type 1. Curr Biol 2005;15:1961–1967.
2. Dölen G, Osterweil E, Rao BSS, et al. Correction of fragile X syndrome in mice. Neuron 2007;56:955–962.
3. Lee YS, Ehninger D, Zhou M, et al. Mechanism and treatment for learning and memory deficits in mouse models of Noonan syndrome. Nature Neurosci 2014;17:1736–1743.
4. van der Vaart T, Overwater IE, Oostenbrink R, et al. Treatment of cognitive deficits in genetic disorders: a systematic review of clinical trials of diet and drug treatments. JAMA Neurol 2015;72:1052–1060.
5. Payne JM, Barton B, Ullrich NJ, et al. A randomized placebo-controlled study of lovastatin in children with neurofibromatosis type 1. Neurology 2016;87:2575–2584.
6. Krab LC, de Goede-Bolder A, Aarsen FK, et al. Effect of simvastatin on cognitive functioning in children with neurofibromatosis type 1: a randomized controlled trial. JAMA 2008;300:287–294.
7. van der Vaart T, Plasschaert E, Rietman AB, et al. Simvastatin for cognitive deficits and behavioural problems in patients with neurofibromatosis type 1 (NF1-SIMCODA): a randomised, placebo-controlled trial. Lancet Neurol 2013;12:1076–1083.
8. Hyman SL, Shores EA, North KN. The nature and frequency of cognitive deficits in children with neurofibromatosis type 1. Neurology 2005;65:1037–1044.
9. Brown JA, Emnett RJ, White CR, et al. Reduced striatal dopamine underlies the attention system dysfunction in neurofibromatosis-1 mutant mice. Hum Mol Genet 2010;19:4515–4528.
10. Molosh AI, Johnson PL, Spence JP, et al. Social learning and amygdala disruptions in Nf1 mice are rescued by blocking p21-activated kinase. Nature Neurosci 2014;17:1583–1590.
11. Omrani A, van der Vaart T, Mientjes E, et al. HCN channels are a novel therapeutic target for cognitive dysfunction in Neurofibromatosis type 1. Mol Psychiatry 2015;20:1311–1321.
12. Mautner VF, Kluwe L, Thakker SD, Leark RA. Treatment of ADHD in neurofibromatosis type 1. Dev Med Child Neurol 2002;44:164–170.
13. Acosta MT, Kardel PG, Walsh KS, et al. Lovastatin as treatment for neurocognitive deficits in neurofibromatosis type 1: phase I Study. Pediatr Neurol 2011;45:241–245.
14. Stivaros S, Garg S, Tziraki M, et al. Randomised controlled trial of simvastatin treatment for autism in young children with neurofibromatosis type 1 (SANTA). Mol Autism 2018;9:12.
15. Bearden CE, Hellemann GS, Rosser T, et al. A randomized placebo-controlled lovastatin trial for neurobehavioral function in neurofibromatosis I. Ann Clin Transl Neurol 2016;3:266–279.

16. Lion-Francois L, Gueyffier F, Mercier C, et al. The effect of methylphenidate on neurofibromatosis type 1: a randomised, double-blind, placebo-controlled, crossover trial. Orphanet J Rare Dis 2014;9:142.
17. Payne JM. Bridging the gap between mouse behavior and human cognition in neurofibromatosis type 1. EBioMedicine 2015;2:1290–1291.
18. Wegscheid ML, Anastasaki C, Gutmann DH. Human stem cell modeling in neurofibromatosis type 1 (NF1). Exp Neurol 2018;299(Pt B):270–280.
19. Pound P, Ebrahim S, Sandercock P, et al.;Reviewing Animal Trials Systematically G. Where is the evidence that animal research benefits humans? BMJ 2004;328:514–517.
20. van der Vaart T, Rietman AB, Plasschaert E, et al. Behavioral and cognitive outcomes for clinical trials in children with neurofibromatosis type 1. Neurology 2016;86:154–160.
21. Neurofibromatosis. Conference Statement. National Institutes of Health Consensus Development Conference. Arch Neurol 1988;45:575–578.
22. Payne JM, Barton B, Shores EA, North KN. Paired associate learning in children with neurofibromatosis type 1: implications for clinical trials. J Neurol 2013;260:214–220.
23. Manly T, Anderson V, Nimmo-Smith I, et al. The differential assessment of children's attention: the Test of Everyday Attention for Children (TEA-Ch), normative sample and ADHD performance. J Child Psychol Psychiatry 2001;42:1065–1081.
24. Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Hillsdale: Lawrence Erlbaum, 1988.
25. Crocker LM, Algina J. Introduction to classical and modern test theory. Cengage Learning, 2006.
26. Lohr KN. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res 2002;11:193–205.
27. Payne JM, Hyman SL, Shores EA, North KN. Assessment of executive function and attention in children with neurofibromatosis type 1: relationships between cognitive measures and real-world behavior. Child Neuropsychol 2011;17:313–329.
28. Lowe C, Rabbitt P. Test/re-test reliability of the CANTAB and ISPOCD neuropsychological batteries: theoretical and practical issues. Cambridge Neuropsychological Test Automated Battery. International Study of Post-Operative Cognitive Dysfunction. Neuropsychologia 1998;36:915–923.
29. Luciana M, Nelson CA. Assessment of neuropsychological function through use of the Cambridge Neuropsychological Testing Automated Battery: performance in 4- to 12-year-old children. Dev Neuropsychol 2002;22:595–624.
30. Bollen KA. Structural equations with latent variables. New York: Wiley, 1989.
31. Pride NA, Korgaonkar M, North KN, et al. The neural basis of deficient response inhibition in children with neurofibromatosis type 1: evidence from a functional MRI study. Cortex 2017;93:1–11.
32. United States Food and Drug Administration. Code of Federal Regulations, Title 21, Volume 5 (21CFR314.126). 2019. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=314.126.
33. Walsh KS, Janusz J, Wolters PL, et al. Neurocognitive outcomes in neurofibromatosis clinical trials: recommendations for the domain of attention. Neurology 2016;87:S21–S30.
34. Finniss DG, Kaptchuk TJ, Miller F, Benedetti F. Biological, clinical, and ethical advances of placebo effects. Lancet 2010;375:686–695.

Supporting Information

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

Table S1. Theoretical justifications for the confirmatory factor analysis model.
Table S2. Confirmatory factor analysis fit statistics.