The Practical Constraints of Clinical Trials

Philip D. Harvey, PhD
Emory University School of Medicine
Atlanta, GA
Potential Conflicts of Interest

Consulting/Advisory Board:

Abbott, Astra-Zeneca, Bristol Myers Squibb, Cephalon, Dainippon Sumitomo, Eli Lilly Laboratories, Johnson & Johnson, Lundbeck/ Solvay/ Wyeth, Memory, Merck, Pfizer, Saegis, Sanofi/Aventis,

Research Funding:

Astra-Zeneca, Bristol-Myers Squibb, Johnson and Johnson, Pfizer
Potential Conflicts of Interest

Royalties Received:

Brief Assessment of Cognition (BACS)
MATRICS Battery (BACS Symbol Coding)
Likely Characteristics of Multisite Clinical Trials

♦ Variability in data quality
  • Site feasibility evaluation required
    • Past history with task or similar
    • Evidence that current situation is adequate
  • Tester screening required
    • What are the minimal educational and experience requirements for the specific task?
Humans will demonstrate their worst characteristics when confronted with the prospect of establishing inter-rater reliability.

Sponsors
- Have had tremendous difficulty organizing training based upon the primary outcome measure (e.g., 6 hours of PANSS training; 2 hours of cognitive training for a cognition trial).
- Companies have repeatedly reverted to old strategies for confronting these new challenges.

PI’s
- Will delay the identification of testers indefinitely if not forced to choose (e.g., on the plane to the investigator meeting), and will change their minds frequently if they are unaware of the training process.

Testers
- Will frequently exaggerate their experience to their PI’s and the central coordinating center.
- Will claim to have performed necessary pre-meeting training.
- Will claim that they were not told they needed to be prepared.
- Will try to get the sponsor to undermine the certification process.

Professional clinical-trials sites may not have sufficient expertise to run sophisticated cognitive neuroscience tasks, however some of them are eager and easily coached, and have large samples.

Academic sites will have fewer patients, slower IRB processes, and greater expertise, which cannot be a problem. Rowers all need to row the same way. Being ‘better’ is worse for the reliability of the group. There is no “I” in inter-rater reliability. OK, there are four, but none of them are capitalized.
How to Love Those You Cannot Trust

♦ Sponsors
  • Give sponsors ample warning of the training and certification processes that will be required for establishing reliability
  • Be prepared to teach them what reliability is and why low reliability is bad (costs money)
  • Have a budget ready
  • Be prepared to cancel involvement in a trial if procedures for adequate reliability are not established. I have wasted so many months following compromised reliability procedures that resulted in useless data.

♦ PI’s
  • Identify testers early
  • Once testers are chosen, they should be changed only in extreme circumstances
  • Keep PI’s aware of training process so that they do not inadvertently sabotage it

♦ Testers
  • Screen them prior to allowing them to be identified as the site tester
  • Training materials must be prepared and sent to sites well ahead of investigators meeting (one month is ideal) to give ample time for inexperienced testers to learn and practice
  • Have a personal “look them in the eye” telephone CONVERSATION with each potential tester: “I will be certifying you at the meeting, so be sure that you are prepared to show me that you are doing the tests correctly.” Be clear that procedures will be certification and not “getting to know you”.
  • Central coordinating center MUST evaluate performance and have power to reject testers
Data Review Processes

♦ The larger the number of sites, the less any individual site feels responsible for assuring the quality of the data (don’t let Kitty Genovese die again!)

♦ Large multi-site studies may require ongoing data quality review for every data point

♦ Catastrophic continuing inaccuracies can be avoided by reviewing data immediately after test administration

♦ Ongoing conference calls for testers to describe problems and how they responded to them are tremendously helpful

♦ Smaller studies with select sites that have academic reputations to lose may require less review from a central site, but it is important to have ongoing checks that data quality procedures are adequate at each site. Random test review from central site keeps everyone honest. Remember that we academic researchers only pretend to live at the top of Maslow’s moral pyramid. Accountability is necessary.
Practical Constraints Will Require Decisions About Task Duration and Complexity

♦ Can the construct be measured in a simpler, cheaper, more parsimonious manner?
♦ Is the cognitive construct really as specific as assumed?
♦ Easily administered tasks have reduced confusion among testers, patients, PIs and readers
♦ Does increased task complexity lead to greater missing data?
♦ Does the size or specificity of the effect outweigh the practical constraints, both financial and scientific?
Responding to Inevitable Crises

♦ Low enrollment leads to additional sites
  • Need to have plan for mobile evaluation of new sites and testers

♦ Personnel changes
  • Adding testers to a site with experience can be less demanding than to a new site
    • Deputizing certified testers as trainers
    • Final approval of testers should be central

♦ Some sites may not be capable of participating
  • “Nothing gets people’s attention like a hanging” – Joe McEvoy
CATIE Schizophrenia Trial Strongman
Lessons from Clinically-Oriented Cognitive Batteries

♦ Shorter is better if it tests the same thing with equal reliability

♦ This may be a big ‘if’ for cognitive neuroscience tasks, but cognitive neuroscientists might not be accustomed to sacrificing theory for brevity and practicality, which clinical trials require
CATIE Schizophrenia Neurocognitive Test Battery

- Wide Range Achievement Test (WRAT-III)—Reading Test (1 visit only)
- Verbal Fluency
  - Controlled Oral Word Association Test (COWAT)
  - Category Instances
- Wechsler Intelligence Scale for Children (WISC-III) Mazes
- Hopkins Verbal Learning Test (HVLT)
- Facial Emotion Discrimination Task (FEDT)
CATIE Schizophrenia Neurocognitive Test Battery (cont.)

- Revised Wechsler Adult Intelligence Scale (WAIS-R) Digit Symbol Test
- Letter-Number Test of Auditory Working Memory
- Grooved Pegboard
- Continuous Performance Test (CPT), Identical Pairs Versions (2, 3, and 4 digits)
- Computerised Visuospatial Working Memory Test
- Computerised WCST
### Stepwise Multiple Regression Predicting Unweighted Mean of Variables in CATIE

<table>
<thead>
<tr>
<th>Variable Entry Based</th>
<th>Total R²</th>
<th>R² Change</th>
<th>Est. time</th>
</tr>
</thead>
<tbody>
<tr>
<td>On Administration Time</td>
<td>R²</td>
<td>Change</td>
<td></td>
</tr>
<tr>
<td><strong>WAIS-R Digit Symbol</strong></td>
<td>.610</td>
<td>.610</td>
<td>3.1</td>
</tr>
<tr>
<td><strong>HVLT Verbal Memory</strong></td>
<td>.722</td>
<td>.112</td>
<td>4.1</td>
</tr>
<tr>
<td><strong>Grooved Pegboard</strong></td>
<td>.790</td>
<td>.068</td>
<td>5.0</td>
</tr>
<tr>
<td><strong>U. Maryland LN Seq</strong></td>
<td>.868</td>
<td>.078</td>
<td>5.9</td>
</tr>
<tr>
<td><strong>Verbal Fluency</strong></td>
<td>.889</td>
<td>.021</td>
<td>8.0</td>
</tr>
<tr>
<td><strong>WISC-R Mazes</strong></td>
<td>.935</td>
<td>.046</td>
<td>11.2</td>
</tr>
<tr>
<td><strong>CPT-IP</strong></td>
<td>.957</td>
<td>.022</td>
<td>13.4</td>
</tr>
<tr>
<td><strong>Visuospatial WM Test</strong></td>
<td>.978</td>
<td>.022</td>
<td>16.2</td>
</tr>
<tr>
<td><strong>WCST-64</strong></td>
<td>1.000</td>
<td>.022</td>
<td>more</td>
</tr>
</tbody>
</table>

F-statistic for all steps was greater than 193.0; all P-values < .0001; N=1035

WAIS-R=Wechsler Adult Intelligence Test, Revised; HVLT=Hopkins Verbal Learning Test; WISC-III=Wechsler Intelligence Test for Children, 3rd ed; WCST=Wisconsin Card Sorting Test

Keefe et al, *Neuropsychopharmacology*, 2006
Baseline Completion Rates

Percentage of Cases

Keefe et al., *Neuropsychopharmacology*, 2006; total n=1427
First Reassessment Completion Rates

Percentage of Cases

Keefe et al., 2007; total n=967
Second Reassessment Completion Rates

Percentage of Cases

Fluency: 99
HVLT: 99
Digit Symbol: 98
LNS: 98
Pegboard: 97
Mazes: 97
WCST: 97
SWM: 90
CPT: 91

Keefe et al., 2007; total n=825
Preliminary Conclusions

♦ Computerized tests look like they have considerably lower completion rates.

♦ By the time you get to assessment 3, you have 40 cases who are not evaluable (because of current or prior missing data) on the HVLT, 407 on the CPT and 336 on the WCST.
BACS and Standard Battery Composite Scores for Patients and Controls

Efficiency of Cognitive Measures in First Episode Psychosis Patients (N=216)

CATIE and BACS scores correlated 0.84 at baseline, 0.90 after one year of treatment.

Both the CATIE and Brief Assessment of Cognition (BACS) batteries had similar levels of information efficiency in aggregate (CATIE: $R^2=.736$, $F=98.68$, df=6,212, $p<.001$; BACS: $R^2=.760$, $F=169.12$, df=4,214, $p<.001$), yet the BACS achieved this in a much shorter period of time.
And it is much more complicated than that

- Prediction of things other than global NP score may result in different subsets of abbreviated assessments
- Prediction of Global NP Score, General Deficit Score (GDS), UPSA Score, SSPA Score, Every day outcomes
- N=246
<table>
<thead>
<tr>
<th>Outcome</th>
<th>Predictors</th>
<th>% Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global NP</td>
<td>TMT B, RAVLT</td>
<td>64%</td>
</tr>
<tr>
<td>GDS</td>
<td>TMTB, Digit Sym</td>
<td>72%</td>
</tr>
<tr>
<td>UPSA</td>
<td>Digit Sym, Bos Name</td>
<td>40%</td>
</tr>
<tr>
<td>SSPA</td>
<td>DS Backward, TMT A</td>
<td>32%</td>
</tr>
<tr>
<td>SLOF Functional</td>
<td>TMT A, Animal FL</td>
<td>24%</td>
</tr>
</tbody>
</table>

Harvey et al., submitted
Structural Equation Modeling Analyses on CATIE Baseline Data (N=1331)

- Null model failed to fit the data
- A unifactorial model based on the nine tests was an improvement in fit; $\chi^2(27)=192.18$, $p<0.001$; CFI=.94, GFI=.97, RMSEA=.077
- A unifactorial model including the five pre-defined domain scores was a considerable improvement in fit over the unifactorial model from the nine tests ($\chi^2(22)=152.27$, $p<.001$; CFI=.98, GFI=.97, RMSEA=.080).
- A five-factor model that included the tests from each of the five cognitive domains as separate factors was a significantly poorer fit compared to the unifactorial model from the five pre-defined domain scores $\chi^2(14)=78.04$, $p<.001$.

Keefe et al, *Neuropsychopharmacology*, 2006
Change in Neurocognitive Composite Score After 2 Months of Treatment

N above histogram. No significant differences between treatments (p=.20). TD=tardive dyskinesia.
Change in Neurocognitive Domains After 2 Months of Treatment

No significant differences between groups (all p>.08)

Perphenazine
Risperidone
Quetiapine
Olanzapine
Ziprasidone

Z-Score Change from Baseline to 2 Months Excluding TD Patients

Processing Speed
Reasoning
Working Memory
Verbal Memory
Vigilance

TD=tardive dyskinesia.
N above histogram. TD=tardive dyskinesia. Overall differences between treatments (p<.05).
<table>
<thead>
<tr>
<th>Measure</th>
<th>Baseline</th>
<th>Week 4</th>
<th>Week 8</th>
<th>Week 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>MATRICS battery</td>
<td>Org/MK</td>
<td>Org/MK</td>
<td>Org/MK</td>
<td></td>
</tr>
<tr>
<td>Ancillary Cognitive Measures</td>
<td>Org/MK</td>
<td>Org/MK</td>
<td>Org/MK</td>
<td></td>
</tr>
<tr>
<td>SCoRS</td>
<td>Org/MK</td>
<td>Org/MK</td>
<td>Org/MK</td>
<td>Org</td>
</tr>
<tr>
<td>UPSA</td>
<td>Org/MK</td>
<td>Org/MK</td>
<td>Org/MK</td>
<td></td>
</tr>
</tbody>
</table>

Org = Organon 24448 (faramptor) Study; MK = Merck 0777; SCoRS = Schizophrenia Cognition Rating Scale; UPSA = UCSD Performance-based Skills Assessment
Collaborators

**Duke**
- Joe McEvoy
- William Wilson
- Trina Walker
- Kirsten Hawkins
- Mike Kraus
- Courtney Kennel
- Miriam McKenzie
- Leslie Yusko

**CATIE NAG**
- Robert Bilder
- Philip Harvey
- Barton Palmer
- Sonia Davis
- Richard Mohs
- James Gold
- Michael Green
- Herbert Meltzer
- *CATIE Neurocognitive Working Group*

**Neurocog Trials, Inc.**
- Kolleen Hurley
- Trina Walker
- Nicole Turcotte
- Rebecca Williams
- Mike Dowling
- Sue Redmond

**CATIE Executive Committee**
- Jeffrey Lieberman
- Joseph McEvoy
- Scott Stroup
- Robert Rosenheck
- Marvin Swartz
- Diana Perkins