Published in final edited form as:

Alzheimers Dement. 2008 November ; 4(6): 390–394. doi:10.1016/j.jalz.2008.09.002.

Comment on administration and scoring of the Neuropsychiatric Inventory (NPI) in clinical trials

Donald J. Connor, PhD, PhD^{a,*}, Marwan N. Sabbagh, MD^a, and Jeffrey L. Cummings, MD^b

^aCleo Roberts Center for Clinical Research, Sun Health Research Institute, Sun City, AZ, 85351, USA

^bDepartment of Neurology, David Geffen School of Medicine at UCLA, Los Angeles, CA, 90095, USA

Abstract

Background—The Neuropsychiatric Inventory (NPI) is commonly used in dementia trials to quantify and qualitate changes in psychiatric symptoms.

Methods—A questionnaire was administered to clinical trial raters to assess whether they were being trained to administer and score the NPI differently between clinical trial protocols.

Results—Responses to the survey indicate that there are differences between clinical trials protocols in how the instrument is administered and scored.

Discussion—Clarification of administration and scoring rules are provided, including the behavioral sampling period, if pre-morbid characteristics are considered, and what behaviors are considered in rating frequency, severity and caregiver distress.

Keywords

Neuropsychiatric Inventory; NPI; dementia; Alzheimer's; clinical trial; administration; procedure

1. Introduction

The Neuropsychiatric Inventory (NPI) is designed to detect, quantify and track changes of psychiatric symptoms in a demented population [1]. It has been used in dementia trials to quantify and qualitate changes during the treatment period [2–5]. It uses a structured, caregiver-based interview format to assess 10 behavioral domains (Delusions, Hallucinations, Agitation, Dysphoria, Anxiety, Apathy, Irritability, Euphoria, Disinhibition, and Aberrant motor behavior). Two additional domains (Nighttime behavioral disturbance, and Appetite/weight changes) are often added [6]. The presence of problematic behaviors in each domain is assessed by asking an informant a screening question followed by a series of yes/no questions. The caregiver or surrogate reporter is then asked to rate the frequency of occurrence of that domain of behaviors (Table 1). The same behaviors are then rated on level of severity (1 = mild, 2 = moderate, 3 = severe). The domain total score is the product of the frequency score multiplied by the severity score for that behavioral domain. A NPI total score is obtained by summing all the individual domain total scores. A measure of the level of caregiver distress is also given (Table 1), but is not included in the NPI total score [7]. The NPI has shown good content

*Corresponding author: Donald J. Connor, Ph.D, Ph.D., Cleo Roberts Center for Clinical Research, Sun Health Research Institute, 10515 West Santa Fe Dr., Sun City, AZ 85351, Tel: 623-875-6500; Fax: 623-875-6504, donald.connor@sunhealth.org.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

validity, concurrent validity, inter-rater reliability (93.6% to 100% for different behaviors) and 3 week test-retest reliability (correlation = 0.79 for frequency and 0.86 for severity ratings) [6].

While the instrument is well standardized and training tapes are available, there may be some variability in how the instrument is administered and scored in clinical trials. Unfortunately this can change the instrument's psychometric properties in addition to introducing measurement biases that may hinder direct comparison of the outcomes between clinical trials. Intra-trial decreases in reliability of the instrument may also occur if the raters are trained in different – and conflicting – ways of administering the scale. A survey of NPI raters at a clinical trials meeting was done to assess whether there are variations in how raters are being trained on this instrument. Some of the more common key issues, variations and questions are briefly addressed.

2. Methods

An eight-item multiple choice survey was distributed to NPI raters at a clinical trials meeting. The survey was voluntary and completed before the NPI training session. The survey was constructed to assess if the rater had been trained to administer or score the NPI differently for different protocols. For each item, the survey asked each rater to, "Please circle all the ways you have been trained to do it or seen it done." Multiple responses to the same item indicated that the rater had been trained (or seen it done) differently in different protocols.

3. Results

Thirteen raters completed the survey. Rater experience with the NPI ranged from 4 – 12 years (7.4 +/- 3.1) and 50 – 500 administrations (196 +/- 153). Forty-six percent of the respondents had experience with both pharmaceutical sponsored studies and those through the Alzheimer's Disease Cooperative Study (ADCS); 31 % had only worked on pharmaceutical company studies; and 23% had only worked on ADCS studies.

Survey results are shown in Table 2. Briefly, reports of variation in training ranged from 85% regarding whether the behavior needs to be a change from previous functioning, to 15% for how caregiver distress is rated.

4. Discussion

4.1 NPI Survey

The NPI survey results indicate that there is some variability in how raters are being trained to administer and score the instrument in different clinical trials. Since raters often are engaged in multiple studies at the same time, having the same instrument administered and scored differently between protocols has the potential to reduce reliability and increase variance through interference effects, especially in longitudinal studies. The variability in procedures also threatens the validity of comparisons of the outcomes between different trials and the use of meta-analytic procedures to combine trial data.

While our study was meant to assess intra-rater variability and not inter-rater reliability, the survey did have some intriguing secondary findings that suggest inter-rater issues may also be of concern. For example, although only 15% of respondents indicated they had been trained in multiple ways of rating distress to the caregiver, of those who endorsed only one response 39% would rate it at the time of the behavior(s); 23% would average the distress rating over the sampling period; and 23% indicated this was not addressed in their training.

While the conclusions from this study are limited by the small sample size (n=13) and the brevity of the survey, they do suggest that some clarification of the administration and scoring methods of the NPI would be beneficial. To that end the authors have addressed some of these issues in the following section.

4.2 Administration and scoring procedures

4.2.1 Behavior sampling interval—The range of time or interval chosen as the behavioral sampling interval varies widely. While most studies appear to use a period of 4 weeks preceding the current interview, some potential variations include use of a 6 week interval; the interval since the last visit to the clinic (may vary throughout the study); the period since the baseline visit; and using different intervals at baseline and follow-up (e.g., baseline visit using an interval from when the disease was diagnosed to present, and other visits using last 4 weeks).

Recommendation: It is suggested that a common sampling interval be used at each visit and sampling intervals should not overlap (e.g. the interval should be no longer than the time to the previous NPI). The validity studies of the NPI were based on a four week interval. The recommendation of 4 weeks should provide a sufficient sampling interval while also being within a period easily recalled by the informant.

4.2.2 Instance vs New Occurrence—The informant may sometimes become confused as to whether the behavior has to merely have been present in the designated interval, or if it has to be a new behavior that occurred in that period. For example, if the patient has shown instances of tearfulness in the last 4 weeks (sampling interval) but was also showing instances of tearfulness 6 weeks ago, should it be counted?

Recommendation: It should be emphasized that the behavior does not have to be a new one, but it does have to have been present at least once in the sampling interval. While the questions should be read as written, if the examiner believes the informant is confused or hesitant a restatement of the question may be given emphasizing **any** instance of that behavioral domain in the last 4 weeks (e.g. “So in the past 4 weeks have there been any instances where your husband ...”).

4.2.3 Present vs Change from Normal Behavior (Anchor point)—A second area of variation is whether the informant is asked to just report if the behavior occurred, or if they are required to rate if the present behavior is a change from some baseline or anchor point. Often this will be to compare the patient’s “normal” behavior before the dementia began to their behaviors in the sampling period. However, some trials may ask that the comparison point be the baseline visit of the trial; some may ask them to compare from the last visit; and some may not require a comparison (e.g. rate current behaviors regardless if they were a preexisting condition). A problem with using the baseline visit as the anchor is that the time interval increases as the trial goes on, and the informant may not remember specifically how the person was at the beginning of the trial. Similarly, comparing to the last visit may give a variable interval, requires them to remember behaviors that occurred in a fairly narrow period of time (4 week period 3 months ago) and may unnecessarily complicate data analysis. Conversely, in a clinical trial it may be argued that requiring a comparison to an anchor point is an unnecessary burden to impose on the informant since we wish to analyze change during the trial. Whether or not the behaviors existed in some form before the onset of the dementia may add little value to the change score analysis and adds another factor that may decrease the reliability of the measure. However, having no baseline comparison may attribute long standing behaviors (e.g. impatience, irritability, mild apathy) to the disease. Factoring out long-standing traits not due to the disease may therefore increase the validity of the measure.

Recommendation: It is suggested that the standard reference point of “compared to how the patient was before he developed dementia” has the advantage of referring to the person’s more general state as the informant knew them and provides a more stable reference point than other anchors such as previous visits.

To help the informant track the requirements of the sampling period, behavioral occurrence and change, a short example given before the test begins may help: “Let me give you an example, if your husband showed a couple instances of impatience in the last 4 weeks, but it was no different than he has been his whole life, it would not be counted. However, if the degree of impatience was different compared to how he was before he developed the dementia then it would be counted. Remember, it doesn’t have to be a new behavior, there just has to have been an instance of it in the last 4 weeks that’s different than how he used to be before the dementia.”

4.2.4 Rating frequency and distress—The NPI rates the behaviors in each domain as to frequency of occurrence, severity of the behavior (to the patient) and level of distress to the caregiver (Table 1). In rating the frequency and severity of the symptoms that are identified, two main questions arise. The first question is whether to rate only the most severe symptom present, or to rate on all symptoms. For example, in the Depression Domain, if the informant reports that their spouse spoke of wanting to die once in the 4 week period but had frequent instances where they appeared sad and would cry several times a week, would the frequency and severity be rated only on the most severe symptom (suicidal ideation) or on all the symptoms mentioned? Rating only the most severe symptom has the problem of asking the examiner to determine which is the most severe behavior as several may be equally severe, and while the most severe behaviors may be infrequent, behaviors of less severity may be of greater frequency and thus have a greater impact on the patient and caregiver.

Recommendation: It is suggested that at the end of the domain subquestions and all information on that domain has been obtained, the examiner repeat back **all** of the endorsed items and ask the informant to rate severity based on all the symptoms present.

If as suggested above, the rating is done on all identified behaviors in the domain, then the question arises whether to rate frequency on the sum of all instances of the behaviors, or on the “average frequency” of the items. Similarly, should the severity rating be on the total severity of all behaviors in the domain, or on the average of the behaviors? For example, if the informant says, “It was marked severity when he said he wanted to die and I couldn’t get him off it for a couple hours, but the other times he was just a little sad and easily cheered up.”

Recommendation: It is suggested that the informant be asked to rate frequency taking into account all of the behaviors in the domain, and rate severity based on all the behaviors. The examiner may say, “So you said that he (list symptoms), on the whole, how often have these behaviors occurred over the last 4 weeks?” For severity say, “So taking all these behaviors into account, on the whole how severe would you say they are?”

Similarly, in rating distress to the caregiver should the level of distress reflect their feelings at the time of the incident or more generally over the sampling period? For example, if the patient only threatened suicide once and it resolved quickly, and they had no other depressive symptoms, the level of distress at the time of the incident may be severe, but the distress caused by that symptom over the full sampling period may only be minimal or none.

Recommendation: In keeping with the ratings of frequency and severity, it is suggested that the examiner ask the informant to rate their general level of distress over the entire sampling period, taking all incidents of that behavioral domain into account.

In rating frequency, severity and caregiver distress a sheet of paper with definitions of each can be given to the patient (Table 1). This helps the interview go more smoothly and increases reliability by better operationalizing the rating levels. Finally, it is always helpful to reemphasize the parameters when repeating back the subquestions that the informant identified as problematic. For example, using the phrase, “So in the last 4 weeks, you noticed they (list behaviors), and this was different than how they were before they were diagnosed with Alzheimer’s disease?”, reinforces the sampling period and that the behaviors were not present before the onset of the disease.

4.2.5 Appetite / eating—One of the additional domains sometimes included in the NPI is Appetite/eating changes. Informants may become confused by this as to whether a weight change has to be over the last 4 weeks or is from their pre-diagnosed state. Also, in this domain the informant occasionally reports a weight change, but no change in appetite. This is difficult to rate on the forms since they require an appetite change accompany any weight change.

Recommendation: It is suggested that the informant be asked about the patient’s weight and appetite over the last 4 weeks compared to how they were before the diagnosis of dementia. If there is weight gain (or loss) but no change in appetite, do not score this item as it is anchored in the behavior of appetite. Do not rate if due to purposeful weight loss (e.g. diet and exercise).

4.2.6 Source of behaviors—In certain circumstances there may be acute behavioral changes due to factors not related to the disease (family member dies, pain from recent accident, illness, etc.). While the change may be salient, it stems from external events and may not reflect a change in the underlying psychological constructs. Alternatively, there may be cases where the patient’s reaction is out of character for them (e.g. not how they would have reacted before onset of the disease) and therefore would indicate a valid change. In clinical practice there is value in attempting to determine if a patient’s reaction to such events represents a change from their premorbid patterns. However, in clinical trials this level of interpretation may significantly reduce reliability.

Recommendation: It is suggested that behavioral changes be scored if present regardless of whether the examiner thinks it is a normal reaction to external events.

The ability to correctly quantitate the effects of a treatment as well as to compare the effects between treatments in different clinical trials is directly dependent on the validity and reliability of the instruments employed. The NPI has been shown to be both valid and reliable, but only insofar as the instrument’s administration and scoring guidelines are well operationalized and consistently applied. This manuscript suggests that there is some variance in aspects of training of the NPI and sets out a series of clarification guidelines to assist in addressing these issues.

Acknowledgements

This work was supported by grants from the NIH (AG19610); the Alzheimer’s Association (NIRG-04-1159); the state of Arizona (AGR 2007-37, AZPD-0011); and the Fox Foundation (D.C., M.S.)

References

1. Cummings JL, Mega M, Gray K, Rosenberg-Thompson S, Carusi DA, Gornbein J. The Neuropsychiatric Inventory: comprehensive assessment of psychopathology in dementia. *Neurology* 1994;44:2308–2314. [PubMed: 7991117]
2. Barcia D, Giles E, Herraiz M, Morinigo A, Roca M, Rodriguez A. Risperidone in the treatment of psychotic, affective, and behavioral symptoms associated to Alzheimer’s disease. *Actas Esp Psiquiatr* 1999;27:185–190. [PubMed: 10431061]

3. Kaufer D. Beyond the cholinergic hypothesis: the effect of metrifonate and other cholinesterase inhibitors on neuropsychiatric symptoms in Alzheimer's disease. *Dement Geriatr Cogn Disord* 1998;9:8–14. [PubMed: 9718229]
4. Mahlberg R, Walther S, Eichmann U, Tracik F, Kunz D. Effects of rivastigmine on actigraphically monitored motor activity in severe agitation related to Alzheimer's disease: A placebo-controlled pilot study. *Archives of Gerontology and Geriatrics* 2007 Jul–Aug;Vol 45(1):19–26. [PubMed: 16963137]
5. Tariot PN, Solomon PR, Morris JC, Kershaw P, Lilienfeld S, Ding C. A 5-month, randomized, placebo-controlled trial of galantamine in AD. *Neurology* 2000;54:2269–2276. [PubMed: 10881251]
6. Cummings JL. The Neuropsychiatric Inventory: assessing psychopathology in dementia patients. *Neurology* 1997;48:S10–S16. [PubMed: 9153155]
7. Kaufer DI, Cummings JL, Christine D, Bray T, Castellon S, Masterman D, MacMillan A, Ketchel P, DeKosky ST. Assessing the impact of neuropsychiatric symptoms in Alzheimer's disease: the Neuropsychiatric Inventory Caregiver Distress Scale. *J Am Geriatr Soc* 1998;46:210–215. [PubMed: 9475452]

Table 1

Scoring levels with example definitions for the NPI.

<u>Frequency of Occurrence:</u>	
1.	Occasionally – less than once per week.
2.	Often – about once per week.
3.	Frequently – several times per week but less than every day.
4.	Very Frequently – once or more per day.
Severity of Behavior: [†]	
1.	Mild – mildly distressing to the patient and not a major problem
2.	Moderate – distressing to the patient but easily overcome by reassurance.
3.	Marked – distressing to the patient and difficult to redirect or deal with.
Distress to you: [†]	
0.	Not at all
1.	Minimally – rarely distressing, and easily tolerated.
2.	Mildly – occasionally distressing, but not a significant problem.
3.	Moderately – somewhat distressing and problematic, but usually tolerable.
4.	Severely – very distressing, and difficult to cope with.
5.	Very Severely or Extremely - markedly distressing and extremely difficult to cope with.

[†] Some studies provide definitions while others just provide the range labels.

Table 2

Percent of survey respondents who reported variance in training rules between different clinical trial protocols or that the topic was not addressed in the training session.

Item	Variability
The time you have been instructed to use as the behavioral sampling interval includes:	69.2%
Does the behavior have to be an new one or could it have occurred before the sampling period?	46.2%
Are the behaviors counted if they occur in the sampling period, or only if they represent a change from some anchor point (e.g. stare before the dementia)?	84.6%
In rating severity of the behaviors in each domain, what symptom(s) have you been instructed to include?	23.1%
In rating frequency of the behaviors, what symptom(s) have you been instructed to include?	38.5%
In rating distress to the caregiver, have you been instructed to rate their feelings at the time of the incident or more generally over the sampling period?	15.4%
Have you been given any aids or written guides to help define the frequency and severity ranges?	53.8%
In certain circumstances there may be acute behavioral changes due to factors not related to the disease (illness, etc.). What have you been instructed to do?	38.5%