STATISTICAL POWER OF STUDIES EXAMINING THE COGNITIVE EFFECTS OF SUBTHALAMIC NUCLEUS DEEP BRAIN STIMULATION IN PARKINSON’S DISEASE

Steven Paul Woods¹, Julie D. Rippeth², Emily Conover¹, Catherine L. Carey³, Thomas D. Parsons⁴, and Alexander I. Tröster⁵

¹Department of Psychiatry, University of California, San Diego, CA, ²Department of Psychiatry, Walter Reed Army Medical Center, Washington, D.C., ³Department of Psychiatry and Behavioral Sciences, Duke University, NC; Joint Doctoral Program in Clinical Psychology, University of California, San Diego, and San Diego State University, CA, ⁴Department of Psychology, Fuller Theological Seminary, CA; Department of Neurology, University of North Carolina, Chapel Hill, NC, and ⁵Department of Neurology, University of North Carolina, Chapel Hill, NC

It has been argued that neuropsychological studies generally possess adequate statistical power to detect large effect sizes. However, low statistical power is problematic in neuropsychological research involving clinical populations and novel interventions for which available sample sizes are often limited. One notable example of this problem is evident in the literature regarding the cognitive sequelae of deep brain stimulation (DBS) of the subthalamic nucleus (STN) in persons with Parkinson’s disease (PD). In the current review, a post hoc estimate of the statistical power of 30 studies examining cognitive effects of STN DBS in PD revealed adequate power to detect substantial cognitive declines (i.e., very large effect sizes), but surprisingly low estimated power to detect cognitive changes associated with conventionally small, medium, and large effect sizes. Such wide spread Type II error risk in the STN DBS cognitive outcomes literature may affect the clinical decision-making process as concerns the possible risk of postsurgical cognitive morbidity, as well as conceptual inferences to be drawn regarding the role of the STN in higher-level cognitive functions. Statistical and methodological recommendations (e.g., meta-analysis) are offered to enhance the power of current and future studies examining the neuropsychological sequelae of STN DBS in PD.

INTRODUCTION

Despite myriad notable limitations to null hypothesis testing (Cohen, 1994; Donders, 2000), this traditional approach to statistical analysis remains decidedly prevalent in psychological research. As such, consideration of statistical power is critical for investigators and research consumers alike. Statistical power refers to
the likelihood that one will accurately reject a null hypothesis when an effect is present (Cohen, 1988). Power is dependent on four primary factors: (a) the sample size; (b) the critical alpha level (0.05 by convention); (c) the effect size observed (or anticipated) in the population of interest; and (d) the specific statistical procedure being used. As a general rule, power values increase with larger sample sizes, stronger effects, higher critical alpha levels, and the use of tests that control more aspects of error variance. In other words, one is more likely to accurately reject a null hypothesis in a study with a large sample and liberal critical alpha level in which substantial effect sizes are evident (Hallahan & Rosenthal, 1996). The generally accepted convention for adequate power is 0.80 (range = 0, 1), which indicates that there is an 80% probability that the null hypothesis will be rejected when true effects are present (Cohen, 1992). Power values below 0.80 increase one’s risk of committing a Type II error (i.e., not rejecting the null hypothesis when true population differences are present).

Cohen (1988) and a multitude of subsequent prominent investigators (e.g., Wilkinson & the American Psychological Association’s Task Force on Statistical Inference, 1999) have urged behavioral scientists to perform power analyses to determine an appropriate sample size given the particular study design and hypothesized effects. Despite such longstanding recommendations and the increasing availability of resources and tools for its calculation, statistical power is not widely reported in published psychological research (e.g., Rossi, 1990; Sedlmeier & Gigerenzer, 1989). The conspicuous absence of power analyses has prompted numerous investigators over the past 30 years to conduct post hoc power analyses of specific psychological literatures. For example, systematic post hoc power reviews are available for psychotherapy (e.g., Kazantzis, 2000) and rehabilitation counseling (e.g., Kosciulek & Szymanski, 1993) outcomes, health psychology (e.g., Maddock & Rossi, 2001), and projective personality assessment (e.g., Acklin, McDowell, & Orndoff, 1992). By and large, such power reviews reach the same general conclusion: Insufficient power remains a widespread problem in psychological research (Cohen, 1992; Sedlmeier & Gigerenzer, 1989). In fact, the failure to consider power has been proposed as one of the “seven deadly sins” of statistical practice in clinical neuropsychology (Millis, 2003). In a recent systematic review of the neuropsychological literature, Bezeau and Graves (2001) conducted post hoc power analyses of 66 articles from the 1998 and 1999 issues of Journal of Clinical and Experimental Neuropsychology, Journal of the International Neuropsychological Society, and Neuropsychology. Consistent with other recent power analyses performed in the psychological literature (Maddock & Rossi, 2001; Sedlmeier & Gigerenzer, 1989), neuropsychological research generally demonstrated insufficient power to detect small and medium effect sizes. However, the median observed population effect size for the neuropsychological articles was large (Cohen’s $d = 0.91$) and corresponded to an ample median power estimate of 0.93. The authors concluded that neuropsychological research typically addresses larger effect sizes than are documented in general psychological research, which may therefore allow for the use of smaller sample sizes.

Although they provide critical and informative data, statistical power reviews of broad literatures such as provided by Bezeau and Graves (2001) may not effectively generalize to specific populations and/or hypotheses (Rossi, 1990). Statistical power is particularly problematic in neuropsychological studies involving clinical populations that are difficult to recruit and enroll in research protocols. For
example, small samples sizes are endemic to studies involving persons with localized brain lesions, low base rate neurological and medical conditions (e.g., prion diseases), and/or who are undergoing novel treatment protocols (e.g., deep brain stimulation). In fact, interventional studies typically exhibit significantly lower power estimates than non-interventional studies, which is often tolerated given the novelty, potential clinical impact, and repeated-measures designs common to clinical trials (Maddock & Rossi, 2001; Vickers, 2003). Although the ethical and logistical factors underlying the smaller sample sizes in interventional studies are legitimate and difficult to circumvent, the resultant limitations on statistical power are nonetheless challenging. Neuropsychological findings derived from such small population samples are often-times contradictory and vary widely across published studies (e.g., Demakis, 2003), which ultimately diminishes one’s ability to draw coherent clinical and conceptual inferences from the scientific literature (e.g., Cohn & Becker, 2003; Maxwell, 2004).

One notable example of this problem is evident in studies examining the cognitive effects of deep brain stimulation (DBS) of the subthalamic nucleus (STN) in persons with Parkinson’s disease (PD). STN DBS is a functional neurosurgical procedure developed to reduce the cardinal motor symptoms of PD (e.g., akinesia, rigidity, and tremor) in treatment refractory patients. Briefly, the surgery involves the bilateral implantation of high-frequency stimulation, quadripolar electrodes into the STN of persons with PD. The electrodes are subsequently linked to a subcutaneous pulse generator (akin to a cardiac pacemaker) that is implanted in the subclavicular area, which allows for outpatient adjustment of stimulation parameters (i.e., frequency, pulse width, and amplitude) to maximize treatment efficacy (Rizzone et al., 2001). A growing body of literature supports the effectiveness of STN DBS for ameliorating off-motor symptoms and dyskinesias, as well as reducing antiparkinsonian medication dosages (Limousin et al., 1998; Pollak et al., 2002). The exact mechanism by which STN DBS reduces the symptoms of PD is controversial, but the high-frequency stimulation procedure may inhibit neuronal activity (e.g., membrane hyperpolarization) in the STN that, in turn, enhances the functioning of nigrostriatal motor output pathways (see Dostrovsky & Lozano, 2002, for a review).

It has been proposed that STN DBS may minimize the risk of cognitive morbidity relative to other neuroanatomical targets (e.g., globus pallidus internus) and surgical techniques (e.g., lesioning methods) (Van Horn, Schiess, & Soukup, 2001). To this end, a recent qualitative review of 16 published studies provided tentative support for the gross cognitive and neurobehavioral safety of STN DBS in PD (Woods, Fields, & Tröster, 2002). Nevertheless, the median sample size of the STN DBS studies in that review was 10 (range = 1–63, all single-group pretest–posttest designs), which raises the concern that this literature may possess inadequate power to detect significant adverse postsurgical cognitive changes. Limited statistical power—in even a small subset of studies—might also falsely increase the variability of STN DBS cognitive outcomes (i.e., adequately powered studies report adverse cognitive outcomes, whereas underpowered studies erroneously report no iatrogenic effects thereby resulting in increased variability in the literature). Indeed, inconsistencies persist across this literature regarding the extent and duration of possible changes in episodic memory, attention, and executive functions (e.g., verbal fluency) after STN DBS. For example, several investigators observed postsurgical declines on measures of verbal fluency, attention, and executive functions, perhaps mediated by the
effects of stimulation on neighboring associative and limbic fronto-striato-thalamo-cortical pathways (Woods et al., 2002). Yet the nature and extent of cognitive decrements after STN DBS is controversial as other studies report no change (and even improvement) in these same cognitive ability areas (e.g., Jahanshahi et al., 2000).

Whether STN DBS is associated with incident cognitive impairment is a question of considerable clinical (as well as conceptual) relevance. Research indicates that cognitive impairment—a common feature of PD (see Tröster & Woods, 2003 for a review)—is associated with greater difficulties independently managing one’s instrumental activities of daily living (IADLs) (e.g., Cahn et al., 1998). Thus, even modest declines in executive or memory functions might be burdensome for patients who already evidence mild neuropsychological deficits prior to surgery (n.b., frank dementia is an exclusion criterion for a majority of surgical candidates). Accordingly, if limited statistical power in the STN DBS literature has masked significant postsurgical cognitive declines, such information would likely alter the informed consent process as regards the potential costs and benefits of surgery. In the absence of formal statistical power analyses, however, it is difficult to determine whether inadequate statistical power might have obscured important cognitive risks associated with STN DBS and/or contributed to inconsistent cognitive outcomes in the literature. Therefore, the aim of the present study was to provide a post hoc estimate of the statistical power of the STN DBS cognitive outcomes literature.

**METHOD**

To identify the relevant published articles, key search terms (e.g., subthalamic nucleus, deep brain stimulation, cognitive, etc.) were entered into the PsychINFO, PubMed, and ISI Web of Science electronic databases for the years 1997 to 2004. In addition, references from articles reporting cognitive outcomes of STN DBS were reviewed to identify other papers of interest that may not have been indexed in the aforementioned databases. To be included in the current power review, an article must have used a repeated-measures design and at least one paired-samples group-level statistical analysis (e.g., a paired-samples t-test) to examine the cognitive sequelae of STN DBS in a sample of persons with PD. Studies that used single- and/or mixed comparison-group designs were included. We excluded review articles, single case studies, statement papers, investigations that used only animal subjects, and studies not published in English.

The 30 articles that met study inclusion criteria were reviewed to determine whether power estimates or standardized effect sizes were reported (Alegret et al., 2001; Ardoüin et al., 1999; Berney et al., 2002; Brusa et al., 2001; Burchiel, Anderson, Favre, & Hammerstad, 1999; Daniele et al., 2003; Dujardin, Defebvre, Krystkowiak, Blond, & Destee, 2001; Funkiewiez et al., 2003, 2004; Gironell, Kulisevsky, Fortuny, Garcia-Sanchez, & Pascual-Sedano, 2003; Halbig et al., 2003; Hershey et al., 2004; Hilker et al., 2003; Jahanshahi et al., 2000; Limousin et al., 1998; Lopiano et al., 2002; Moretti et al., 2003; Moro et al., 1999; Morrison et al., 2004; Patel et al., 2003; Perozzo et al., 2001; Pieroff et al., 2000; Saint-Cyr, Trépanier, Rajeev, Lozano, & Lang, 2000; Schneider et al., 2003, Schroeder et al., 2003, 2004; Trépanier et al., 2000; Volkmann et al., 2001; Whelan, Murdoch, Theodoros, Hall, & Silburn, 2003; Witt et al., 2004). The G*Power statistical package (Buchner, Faul, &
Erdfelder, 1997; Erdfelder, Faul, & Buchner, 1996) was then used to calculate the statistical power of each study. Specifically, post hoc power calculations for paired-samples t-tests were generated considering each individual study’s sample size, associated degrees of freedom, and a critical alpha level of 0.05. The effect size index \( f = [\frac{\sigma_m}{\sigma}] \) is recommended for study designs in which \( k \geq 2 \) (Cohen, 1988), as with the STN DBS literature where multiple repeated measures designs are commonplace. Accordingly, power estimates were conducted using a priori defined Cohen’s \( f \) values for small \((f = 0.10)\), medium \((f = 0.25)\), and large \((f = 0.40)\) effect sizes. Cohen’s \( f \) values—which are always positive and range from zero to an indefinite upper limit—are interpreted as the standard deviation of the standardized means in a given set of populations (Cohen, 1988). Following recommendations from Zakzanis (2001) and Rossi (1990), we also calculated power estimates for very large \((f = 1.5)\) Cohen’s \( f \) values since traditional effect size conventions may not adequately cover the range of effects that might be of clinical interest.

In a second analysis, we derived power values specifically for verbal fluency tasks using the observed rather than a priori defined effect sizes. Verbal fluency tasks were reported in 19 (63%) of the 30 STN DBS studies, making them the most commonly employed cognitive measures in this literature. Power values were calculated for these studies using the observed effect size \((f)\), sample size, degrees of freedom, and a critical alpha level of 0.05. We were unable to derive power values for 5 of the 19 verbal fluency studies because they did not report sufficient data to generate an effect size.

RESULTS

None of the 30 studies reported statistical power analyses or formal measures of effect size. The median sample size of persons with PD undergoing STN DBS in these studies was 14 (interquartile range = 8, 22). Descriptive statistics derived from the post hoc power analyses are presented in Table 1. Results revealed overall minimal power for the detection of conventionally small, medium, or large effect sizes (range = 0.05, 0.91). Only 7% \((n = 2)\) of the studies reviewed demonstrated adequate power \((\geq 0.80)\) to detect a traditionally large effect.

Power estimates based on observed effect sizes from the 14 studies that reported sufficient data on verbal fluency are displayed in Table 2. The mean

| Table 1 Estimated power of studies reporting cognitive outcomes of STN DBS in PD \((N = 30)\) |
|-----------------------------------------|----------|----------|----------|----------|----------|
| Effect size \((f)\)                      | \(M\)    | \(SD\)   | Median   | IQR      | Range    |
| Small \((f = .10)\)                     | .07      | .02      | .06      | .05, .07 | .05, .13 |
| Medium \((f = .25)\)                    | .18      | .13      | .13      | .09, .20 | .06, .54 |
| Large \((f = .40)\)                     | .34      | .23      | .25      | .17, .43 | .07, .91 |
| Very large \((f = 1.5)\)               | .94      | .13      | .99      | .95, .99 | .32, 1.00 |

Note. These data reflect post hoc statistical power estimates generated using standard effect size conventions (cf. observed effect sizes), which were adapted from Cohen (1988) and Zakzanis (2001). Cohen’s \( f \) values \([f = (\sigma_m/\sigma)]\) reflect the \( SD \) of the standardized means in a population (Cohen, 1988). DBS = deep brain stimulation; IQR = interquartile range; PD = Parkinson’s disease; STN = subthalamic nucleus.
Cohen’s f effect size of 0.23 (SD = 0.15) in these studies provided a mean observed power of 0.16 (SD = 0.12) to detect postsurgical verbal fluency changes. Not surprisingly, the five studies that reported significant declines in verbal fluency after STN DBS demonstrated superior power ($M = 0.25, SD = 0.09$) to the nine that observed no such changes ($M = 0.11, SD = 0.13$), $X^2 (1, N = 14) = 5.6, p = .02, d = 1.27$, power = 0.55.

**DISCUSSION**

Published studies on the neuropsychological sequelae of STN DBS in PD largely suggest that this procedure is associated with minimal risk of gross cognitive decline for a majority of appropriate surgical candidates. In support of this contention, data from the present review indicate that, on average, studies within the STN DBS literature demonstrate a 94% chance of detecting such substantial postsurgical cognitive declines (i.e., very large effect sizes) if they were truly present. However, it remains uncertain whether STN DBS leads to milder cognitive decrements in attention, verbal memory, and executive functions (see Woods et al., 2002) nevertheless might be of clinical significance. Our review revealed surprisingly low statistical power to identify conventionally small, medium, and large effect sizes in the STN DBS cognitive outcomes literature; for example, the studies reviewed averaged only a 34% probability of accurately detecting the presence of a traditionally large effect. In fact, only two (7%) of the 30 published studies reviewed afforded sufficient power ($\geq 0.80$) to detect a hypothesized large effect size. Low power was evident even when we examined the observed medium effect sizes associated with postsurgical changes in verbal fluency, which was the most commonly assessed domain. Notably, studies that reported significant postsurgical verbal fluency declines displayed superior power to those that observed no effect of DBS on verbal fluency performance.

It is widely held that the substantial gains in motor functioning and health-related quality of life after STN DBS (e.g., Pollak et al., 2002) outweigh the risk of cognitive decline for a large proportion of surgical candidates (Woods et al., 2002). However, evidence for low statistical power to detect small, medium, and large effect sizes precludes one from drawing conclusions regarding the full impact of STN DBS on cognitive functions. This is of considerable importance because Type II error is especially risky when assessing cognitive morbidity associated with STN DBS (cf., an elevated Type I error risk would fall conservatively in the direction of safety). While the presence of Type II error in interventional studies designed to detect the benefits of a given procedure may result in the erroneous

### Table 2

<table>
<thead>
<tr>
<th>Verbal fluency statistic</th>
<th>M</th>
<th>SD</th>
<th>Median</th>
<th>IQR</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect size (f)</td>
<td>.23</td>
<td>.15</td>
<td>.18</td>
<td>.14, .30</td>
<td>.05, .61</td>
</tr>
<tr>
<td>Power</td>
<td>.16</td>
<td>.12</td>
<td>.11</td>
<td>.06, .23</td>
<td>.06, .45</td>
</tr>
</tbody>
</table>

*Note. N = 14; DBS = deep brain stimulation; IQR = interquartile range; PD = Parkinson’s disease; STN = subthalamic nucleus. Cohen’s f values [$f = (\sigma_m)/\sigma$] reflect the SD of the standardized means in a population (Cohen, 1988).*
conclusion that a given treatment is ineffective (Maddock & Rossi, 2001), false negatives in the detection of adverse side effects are potentially more perilous. Indeed, postsurgical cognitive decrements associated with large (and perhaps even medium) effect sizes may adversely impact performance of IADLs for persons with PD (e.g., Cahn et al., 1998; Chaytor & Schmitter-Edgecombe, 2003), especially for patients with mild presurgical cognitive deficits for whom even a slight decrement in neuropsychological performance may lead to IADL complications. Accordingly, the possibility of significant Type II error in the existing STN DBS cognitive outcomes literature might influence the clinical decision-making process regarding the risk–benefit ratio of cognitive morbidity and considerable motor gains associated with this procedure. Surgical candidates and their caregivers should be informed regarding the possible risk of unforeseen cognitive decrements associated with STN DBS. A postsurgical neuropsychological evaluation may be indicated to assess the possible incidence of subtle cognitive, psychiatric, and/or functional impairment, as well as to inform interventions that would maximize adherence to postsurgical medical regimens (Woods et al., 2002).

Inadequate statistical power necessitates cautious interpretation of the conceptually driven investigations of the STN’s involvement in higher-level cognitive functions. Given the relative ease with which stimulation parameters may be manipulated on an outpatient basis, DBS provides the cognitive neuropsychologist a unique opportunity to employ more rigorous, hypothesis-driven experimental methodologies. In response, emerging studies are exploring the nature and extent of the STN’s role in specific aspects of language, executive functions, and social cognition using dissociation methodologies (e.g., on-off-on stimulation designs) that require acceptance of a true null hypothesis. Nevertheless, absence of evidence cannot be taken as convincing evidence of absence when interpreting the literature regarding the neuropsychological sequelae of STN DBS in persons with PD. As eloquently stated by Cohen (1988):

An analysis which finds that the power was low should lead one to regard the negative results as ambiguous, since failure to reject the null hypothesis cannot have much substantive meaning when, even though the phenomenon exists (to some given degree), the a priori probability of rejecting the null hypothesis was low. (p. 4)

The small samples in the STN DBS literature are ostensibly a function of logistical and ethical problems inherent to research evaluating novel neurosurgical procedures (see Fields & Tröster, 2000). Although the use of repeated-measures methodologies may increase study power by reducing variability parameters (Vickers, 2003), investigators are nevertheless encouraged to maximize the number of enrolled study participants. Such efforts will likely be facilitated by the increasing availability of STN DBS subsequent to its approval by the Food and Drug Administration. Ideally, sample sizes would be dictated by a priori power analyses. Numerous texts (e.g., Cohen, 1988), published articles (e.g., Hallahan & Rosenthal, 1996), and computer software packages (e.g., G*Power: Buchner et al., 1997) are readily accessible in this regard. Sample sizes informed by a priori power analyses will increase statistical rigor, as well as afford investigators the opportunity to utilize more complex statistical procedures to examine possible mediators of postsurgical
A few limitations of the current study should be highlighted. Firstly, not all of the studies included in this review were designed for the primary purpose of evaluating cognitive outcomes. Secondly, the post hoc statistical power analyses reported herein were conducted specifically for paired-samples statistical tests and therefore do not necessarily generalize to other reported statistical analyses (e.g., between-group comparisons or regression-based analyses) (see Rossi, 1990). Thirdly, the apparent variability in the STN DBS cognitive outcomes literature may be partly attributable to factors other than low statistical power. For instance, heterogeneity in participant demographics and disease characteristics, surgical techniques, stimulation parameters, variable test-retest intervals, practice effects, postsurgical medication changes, and/or Type I error due to multiple exploratory statistical comparisons (see Maxwell, 2004) might also contribute to inconsistent findings (see Woods et al., 2002, for review). When appropriate, investigators might therefore consider decreasing sample heterogeneity, using highly reliable and valid dependent measures with continuous outcome variables (cf., dichotomous dependent variables), limiting critical alpha corrections, and pooling data across multiple research centers in an effort to increase power (e.g., Hallahan & Rosenthal, 1996; Maddock & Rossi, 2001).

Finally, meta-analyses are another means of potentially increasing the statistical power of existing literatures that, like STN DBS, are hampered with small sample sizes (e.g., Demakis, 2003). Enhanced statistical power is one of the most commonly cited benefits of meta-analytic studies (Cohn & Becker, 2003). A fundamental aim of a meta-analysis is to estimate a population effect size ($\theta$) by examining findings across independent studies (Demakis, 2006). Meta-analyses can increase statistical power by lowering the standard error associated with the population effect size, which ultimately provides a smaller confidence interval and thereby increases one's power to detect true nonzero population effects (Cohn & Becker, 2003). Meta-analyses would also allow for a more precise and powerful examination of potential moderator variables (e.g., stimulation parameters) that might influence the neuropsychological outcomes of STN DBS. A priori power analyses should also be considered to evaluate the risk of Type II error for meta-analyses, particularly when studies with small sample sizes are involved (Hedges & Pigott, 2001).

ACKNOWLEDGEMENT

The authors thank Sarah Gibson for her assistance in compiling the references.

REFERENCES


