A Responsive Outcome for Parkinson's Disease Neuroprotection Futility Studies

Jordan J. Elm, MA,[1] Christopher G. Goetz, MD,[2] Bernard Ravina, MD,[3] Kathleen Shannon, MD,[2] George Fredrick Wooten, MD,[4] Caroline M. Tanner, MD, PhD,[5] Yuko Y. Palesch, PhD,[1] Peng Huang, PhD,[1] Paulo Guimaraes, PhD,[1] Cornelia Kamp, MBA, CCRC,[6] Barbara C. Tilley, PhD,[1] and Karl Kieburtz, MD, MPH,[6] for the NET-PD Investigators

Futility studies are designed to test new treatments over a short period in a small number of subjects to determine if those treatments are worthy of larger and longer term studies, or if they should be abandoned. An appropriate outcome measure for a neuroprotection futility study in Parkinson's disease (sensitive to tracking disease progression in the short-term) has not been determined. Data sets from three clinical trials were used to compare Parkinson's disease outcome measures. Total Unified Parkinson's Disease Rating Scale (UPDRS; Mentation + Activities of Daily Living + Motor) change and Motor plus Activities of Daily Living UPDRS change, measured in untreated patients, required the smallest sample sizes of all the outcome measures explored. Other outcomes (UPDRS Motor, UPDRS Activities of Daily Living, and time to need levodopa) required somewhat larger sample sizes. Futility designs in Parkinson's disease are feasible in terms of short duration and small sample size requirements, and this design is being applied in two ongoing Parkinson's disease studies to select agents for future larger and longer term neuroprotection studies.

Ann Neurol 2005;57:197–203

Neuroprotection clinical trials for Parkinson's disease (PD) often require large sample sizes and frequently can involve long study participation. A pilot study with a futility design[1] can help avoid unnecessary efficacy studies. The futility design is applicable to phase II trials, and it identifies futile (or noneffective) treatments quickly with a small number of patients.[1] A single treatment arm is compared with a predetermined lower limit of success (or an upper limit of worsening) in a one-sample test. If the treatment is better than or equal to the predetermined limit, it is a candidate for a phase III study. However, if the treatment is worse than the predetermined limit (ie, treatment is futile), then it would be eliminated from further study.

In a futility design, an outcome measure must be sensitive enough to detect PD progression over a short period (ie, 12 months). An ideal short-term outcome measure would be sensitive enough to detect small changes and could discriminate between an effect on symptoms alone and an effect on the progression of PD. Theoretically, this distinction could be accomplished by a biomarker of disease progression or the use of a washout period sufficient to eliminate any pharmacodynamic effect on PD symptoms. Currently, however, there is no well-established in vivo biological marker, and washout periods are limited by practical and ethical concerns regarding leaving patients untreated and uncertainty about the pharmacodynamics of drugs.[2]

Clinical rating scales remain the accepted approach for measuring PD clinical progression and assessing a treatment effect, although they do not easily separate disease-modifying from symptomatic effects. There are several available clinical rating scales for assessing PD impairment and disability. The most commonly used scale is the Unified Parkinson's Disease Rating Scale (UPDRS).[3] Other frequently collected measures include Hoehn and Yahr (H&Y),[4] Schwab and England Activities of Daily Living (SE ADL),[5] Clinical Global Impressions,[6] 12-Item Short-Form Health Survey (SF-12),[7] U.S. Parkinson's Disease Questionnaire (PDQ-39),[8] Mini–Mental State Examination,[9] Beck Depression Index,[10] and Hamilton Depression Index.[11]

The National Institute of Neurological Disorders and Stroke has sponsored a program of phase II and III trials to identify neuroprotective treatments for PD

From the [1]Department of Biostatistics, Bioinformatics and Epidemiology, Medical University of South Carolina, Charleston, SC; [2]Department of Neurological Sciences, Rush-Presbyterian-St. Luke's Medical Center, Chicago, IL; [3]National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD; [4]Department of Neurology, University of Virginia Health Sciences Center, Charlottesville, VA; [5]The Parkinson's Institute, Sunnyvale, CA; and [6]Department of Neurology, University of Rochester, Clinical Trials Coordination Center, Rochester, NY.

Received Jul 26, 2004, and in revised form Oct 12. Accepted for publication Oct 13, 2004.

Published online Jan 26, 2005, in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/ana.20361

Address correspondence to Ms Elm, Medical University of South Carolina, Department of Biostatistics, Bioinformatics & Epidemiology, 135 Cannon St., Suite 303, Charleston, SC 29425.
E-mail: elmj@musc.edu

Published by Wiley-Liss, Inc., through Wiley Subscription Services

(Neuroprotection Exploratory Trials in PD [NET-PD]). Twelve compounds were identified, by an independent group, as possible neuroprotection agents worthy of further study.[12] Futility studies are being used to help select the best candidates for long-term studies. In planning these studies, we examined data sets to answer two study design questions. First, what is the short-term outcome measure that best captures PD progression in 12 months? Second, can short-term change be detected in patients who receive stable treatment with dopaminergic therapy?

Materials and Methods

Data Sets

The Neuroprotection Exploratory Trials in PD Steering Committee sought historical PD data sets from academic investigators and industry. We considered only data sets with untreated patients or patients stably treated with dopaminergic therapy. Several data sets received were not included in this analysis because they lacked sufficient 1-year follow-up data, the exact date of initiation of dopaminergic treatment was not available, or the doses of dopaminergic therapy were variable to control increasing disease-related impairments and disability. The data sets analyzed in this study were previous clinical trials conducted by the Parkinson Study Group, including Deprenyl and Tocopherol Antioxidative Therapy of Parkinsonism (DATATOP),[13] Coenzyme Q10 Evaluation-2 (QE2),[14] and Comparison of the Agonist Pramipexole versus Levodopa on Motor Complications of Parkinson's Disease (CALM-PD).[15] The DATATOP and QE2 trials enrolled patients with recent onset of PD who did not require dopaminergic therapy. In contrast, the CALM-PD trial enrolled patients with early onset requiring their first dopaminergic therapy at the start of the trial.[15] All analyses conducted in this study used only control groups receiving either placebo (QE2: N = 16), placebo plus tocopherol (DATATOP: N = 401), or carbidopa/L-dopa (CALM-PD: N = 150). The starting point for the QE2 and DATATOP studies was the baseline visit. Because CALM-PD started treatment for subjects receiving L-dopa with dosage adjustment over 10 weeks, we chose the 10-week time point when the carbidopa/L-dopa dose had been stabilized as the baseline visit for this analysis.

Scales Available for Study

The following clinical rating scales were available for analysis as potential primary outcome measures for a futility study: Total UPDRS (Mentation + ADL + Motor) score,[3] Motor plus ADL UPDRS score, UPDRS Mental scale,[3] UPDRS ADL scale,[3] UPDRS Motor scale,[3] H&Y,[4] and SE ADL.[5]

Time to Event

The onset of need for dopaminergic therapy (either initial or supplemental) was considered as a possible short-term outcome measure. Notably, in DATATOP and QE2, the onset of need for dopaminergic therapy was of primary interest, whereas in CALM-PD, the decision to administer supplemental L-dopa was of secondary interest.

Time to onset of postural instability was also considered for a short-term outcome by exploring combinations of the UPDRS items Falling, Freezing, and Postural Instability and the H&Y at various cut points. The first occurrence of one of the above was counted as a failure in a survival analysis. More stringent failure definitions (ie, occurrence of postural instability for two consecutive visits) were also considered.

Time Period of Investigation

Because futility studies should be short-term, we analyzed only data up to 12 months.

Statistical Techniques

Each clinical rating score was described across time with box and whisker plots. For each clinical rating score, change from baseline to 6 and 12 months was examined. Last observation carried forward (LOCF) was used to adjust for missing visits. Paired *t* tests were conducted to test whether the change from baseline was significantly different from zero for continuous/normally distributed variables, and the signed-rank test was used for ordinal variables (H&Y and SE ADL). For outcomes that take time to event into account (eg, time to L-dopa), rates were estimated using Kaplan–Meier curves. The one-sided 99% pointwise confidence interval (CI) for the Kaplan–Meier estimate at 12 months was computed and examined to determine if it included zero. For the DATATOP and QE2 data sets, we chose the last score taken at the time when open-label L-dopa was determined necessary and carried this score forward for all subsequent visits to 6 or 12 months. Likewise, in the CALM-PD data, if the decision was made to administer supplemental L-dopa, then scores from this visit were carried forward.

Sample size calculations for a futility study were done to test H_0: (observed change [or proportion] at 12 months $\leq \theta$) versus H_A: (observed change [or proportion] at 12 months $> \theta$), where θ (the maximum acceptable worsening) is defined as 30% less worsening than that observed in historical placebo data. Therefore, under the null hypothesis, we assume that patients given the treatment under study have at least a 30% better outcome than historical control subjects. Thus, for each sample size calculation, θ is considered to be 70% of the historical control value for the outcome measure under consideration (whether change or proportion). For example, if the proportion of historical control subjects having an event at 12 months was 10%, then θ is defined as 7%. If the null hypothesis is rejected, the treatment will be considered futile. Failure to reject the null hypothesis indicates the treatment warrants further evaluation in a phase III study.

Sample size calculations for continuous/ordinal outcomes were based on a one-sample *t* test. Sample size calculations for the time to event outcomes assumed exponential survival, accrual of subjects over a 6-month period, and exactly 12 months of follow-up after end of accrual. All sample sizes were computed to provide at least 85% power to reject the null hypothesis of nonfutility if, in fact, the true mean change was greater than or equal to the average historical control change (or proportion), using a one-tailed test at the 10% level of significance.[16] This level of significance reduces the sample size, whereas maintaining an acceptably low chance of incorrectly calling a drug futile.[16]

Criteria for a Short-term Outcome Measure

Outcome measures of choice were conceptually defined as those that would track disease progression in the short-term and may be correlated with long-term disease progression. Therefore, only outcomes showing short-term changes indicating clinical decline significantly different from baseline ($p < 0.01$) were considered. This α value was selected to identify only measures with a robust potential for detecting clinical changes. For time to event outcomes, the 99% CI must not include zero.

Results

Rating Scales

The Figure shows box plots of Total UPDRS change over time for two different types of samples: (1) patients receiving a stable dose of carbidopa/L-dopa (CALM-PD); and (2) two groups of patients receiving no dopaminergic therapy (DATATOP and QE2). In the first group, the box plots show mean and median changes near zero at each time point (3, 6, 9, and 12 months) and range from −20 to 20. By 12 months, there is only a slight (1.2 point) worsening compared with baseline (week 10 visit). In contrast, both groups of subjects not receiving dopaminergic therapy showed mean worsening over time, noticeable as early as 4 to 6 months and reaching approximately 10 points after 12 months. Box plots for the Motor plus ADL UPDRS, ADL UPDRS, Motor UPDRS, and SE ADL scales showed patterns similar to Total UPDRS (not shown).

The Table shows the change from baseline to 6 and 12 months for the UPDRS scales, H&Y, and SE ADL. For the sample of patients stably treated with L-dopa (CALM-PD), no change scores were significantly different from zero at 6 months ($p > 0.01$) and only Motor plus ADL UPDRS change was significantly different at 12 months ($p < 0.01$). For DATATOP, all outcome measures showed statistically significant deterioration at 6 and 12 months ($p < 0.01$). QE2 data showed declines similar in magnitude to DATATOP at both 6 and 12 months.

Among all outcomes in the same data set, the coefficient of variation for Motor plus ADL UPDRS was the smallest, followed closely by Total UPDRS. For example, in DATATOP, for change at 12 months, the coefficient of variation for Motor plus ADL UPDRS was 106% (compared with 107% for Total, 363% for Mental, 118% for Motor, 117% for ADL, 175% for H&Y, and −120% for SE). Thus, the data scatter compared with the mean was smallest for Motor plus ADL and Total UPDRS at both time points.

Onset of Need for Dopaminergic Therapy

Time until need for dopaminergic therapy, either initial or supplemental, was considered as a short-term survival outcome measure. Approximately 40% of patients in DATATOP (placebo + tocopherol) and QE2 (placebo) met the need for L-dopa therapy within 12 months from baseline (99% CI did not include zero). Similarly, in CALM-PD (L-dopa arm), 28% of patients, after being stably treated with L-dopa, had a need for supplemental L-dopa therapy within 12 months (99% CI did not include zero).

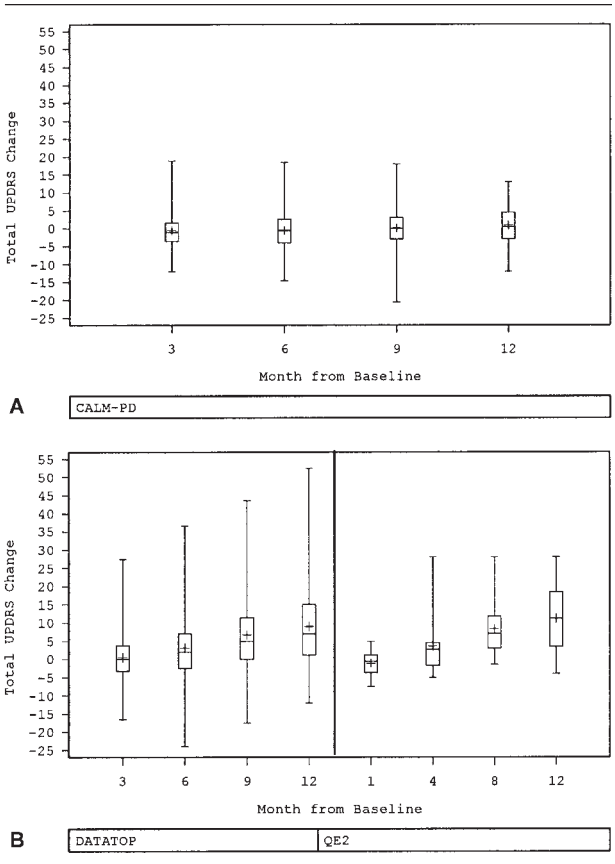

Fig. Total Unified Parkinson's Disease Rating Scale change over time for two different types of Parkinson's disease samples. (A) Subjects stably treated with dopaminergic therapy. Baseline is defined as week 10 visit, at which point the carbidopa/L-dopa doses were stabilized for long-term treatment. If supplemental L-dopa needed to be added, the score at that visit associated with need for dosage increase was carried forward for this analysis. Data from Comparison of the Agonist Pramipexole versus Levodopa on Motor Complications of Parkinson's Disease (CALM-PD; L-dopa arm: N = 135). (B) Subjects not receiving dopaminergic therapy at baseline. If the need for L-dopa was reached before 12 months, then the score at that visit associated with the need for L-dopa was carried forward for this analysis. Data from Deprenyl and Tocopherol Antioxidative Therapy of Parkinsonism (DATATOP; placebo ± tocopherol arms: N = 399) and Coenzyme Q10 Evaluation-2 (QE2) trial (placebo arm: N = 16). The length of the box represents the interquartile range. The plus sign *represents the mean. The* horizontal line *in the box represents the median. Whiskers extend to minimum and maximum values. Missing or out of window values were imputed using the last observation carried forward.*

Table. Mean (SD) Change from Baseline and Coefficient of Variation (%CV) for Various Outcome Measures

		Change[a] from Baseline to Month 6 (±30 days)													
		UPDRS Total		UPDRS Motor + ADL		UPDRS Mental		UPDRS Motor		UPDRS ADL		Hoehn & Yahr		Schwab & England	
Data Set	N	Mean (SD)	%CV	Mean (SD)	%CV	Mean (SD)	%CV	Mean (SD)	%CV	Mean (SD)	%CV	Mean (SD)	%CV	Mean (SD)	%CV
CALM-PD	135	0.01 (5.43)	73,287	−0.01 (5.21)	−70,332	0.01 (0.98)	6,648	−0.04 (4.43)	−9,966	0.03 (2.20)	7,410	−0.04 (0.45)	−1,102	N/A	N/A
QE2	16	3.56 (7.71)	216	3.13 (6.70)	214	0.44 (1.50)	344	0.88 (3.45)	395	2.25 (4.33)	192	0.00 (0.32)	—	−1.88 (4.03)	−215
DATATOP	399	5.37[b] (9.28)	173	5.19[b] (8.83)	170	0.19[b] (1.40)	742	3.47[b] (6.68)	193	1.72[b] (3.23)	188	0.16[b] (0.48)	307	−3.60[b] (6.58)	−183

		Change from Baseline to Month 12 (± 30 days)													
		UPDRS Total		UPDRS Motor+ADL		UPDRS Mental		UPDRS Motor		UPDRS ADL		Hoehn & Yahr		Schwab & England	
Data Set	N	Mean (SD)	%CV	Mean (SD)	%CV	Mean (SD)	%CV	Mean (SD)	%CV	Mean (SD)	%CV	Mean (SD)	%CV	Mean (SD)	%CV
CALM-PD	135	1.20 (5.40)	452	1.17[b] (5.18)	444	0.04 (0.87)	2342	0.65 (4.27)	659	0.50 (2.28)	452	0.02 (0.48)	2,179	N/A	N/A
QE2	16	11.25[b] (9.30)	83	10.38[b] (8.34)	80	0.88 (1.41)	161	6.19[b] (6.15)	99	4.19[b] (4.26)	102	0.09 (0.38)	400	−6.88[b] (6.29)	−92
DATATOP	399	10.11[b] (10.83)	107	9.70[b] (10.29)	106	0.41[b] (1.49)	363	6.45[b] (7.60)	118	3.25[b] (3.81)	117	0.29[b] (0.51)	175	−6.55[b] (7.83)	−120

Data includes only placebo/control arms. All missing values after baseline were imputed using LOCF. The 4-month visit was used as a surrogate for 6 months in QE2 because no visits were collected at 6 months.

[a]Change = X months (±30 days) score - baseline score.
[b]Significantly different from zero ($p < .01$).

SD = standard deviation; UPDRS = Unified Parkinson's Disease Rating Scale; CALM-PD = Comparison of the Agonist Pramipexole versus Levodopa on Motor Complications of Parkinson's Disease; QE2 = Coenzyme Q10 Evaluation-Z; DATATOP = Deprenyl and Tocopherol Antioxidative Therapy of Parkinsonism; N/A = not applicable.

Onset of Postural Instability

When the onset of postural instability was defined as the first visit where either UPDRS Postural Instability was greater than 0, UPDRS Falling was greater than 0, or UPDRS Freezing was greater than 2, then the proportion of subjects that became posturally unstable in 12 months was at its greatest (44% in DATATOP [99% CI did not include zero], 31% in QE2 [99% CI included zero], 23% in CALM [99% CI did not include zero]). However, for all definitions of postural instability explored, more than half of the patients who had an onset of postural instability reverted to normal at a subsequent visit.

When the onset of postural instability was defined more rigorously, with patients required to meet the definition of postural instability at two consecutive visits without reverting to normal, then little short-term change could be detected. Only 8% (99% CI did not include zero) in DATATOP, 0% (99% CI included zero) in QE2, and 1% (99% CI included zero) in CALM-PD met this stringent definition of onset of postural instability over 12 months.

Sample Size Calculations for a Futility Study

Using DATATOP (placebo ± tocopherol) as the historical control, we computed sample sizes for all outcomes: Motor + ADL UPDRS (n = 69), Total UPDRS (n = 70), Motor (n = 84), ADL (n = 84), Mental (n = 830), H&Y (n = 186), SE ADL (n = 87), onset of need for L-dopa (n = 85), and onset of postural instability (for two consecutive visits; n = 577). Thus, Motor + ADL UPDRS and Total UPDRS required the smallest sample sizes, followed closely by ADL UPDRS, Motor UPDRS, and need for L-dopa.

In contrast, using the same parameters, but using the CALM-PD L-dopa data as the historical control, the sample sizes needed are much larger: Motor + ADL UPDRS (n = 1,172), Total UPDRS (n = 1,184), Motor (n = 2,578), ADL (n = 1,243), Mental (n = 28,243), H&Y (n = 34,381), onset of need for additional L-dopa (n = 138), and onset of postural instability (for two consecutive visits; n = 4,867).

Discussion

Clinical Rating Scales

Patients not requiring dopaminergic therapy at baseline showed statistically significant worsening over 12 months in all clinical rating scales, and relatively small sample sizes would be needed to test a futility hypothesis when the maximum acceptable worsening is 70% of the historical control rate (ie, 30% less worsening than historical data). In these untreated patients, the Motor + ADL UPDRS and Total UPDRS (Mentation + ADL + Motor) were the most responsive clinical rating scales to short-term change and required the smallest sample sizes. Change in 6 months in these two outcomes may also be appropriate for short-term studies but would require larger sample sizes.

In contrast, the sample of patients treated with L-dopa in CALM-PD showed no significant changes at 6 or 12 months in any clinical rating measure except

for Motor plus ADL change at 12 months. Because the patient demographics in CALM-PD were similar to the other studies for sex (male patients: 66% in DATATOP; 75% in QE2 placebo; 66% in CALM-PD L-dopa), age (61.1 ± 9.4 years in DATATOP placebo ± tocopherol; 63.1 ± 12.1 years in QE2 placebo; 60.9 ± 10.5 years in CALM-PD L-dopa), and disease duration (1.23 ± 1.1 years in DATATOP placebo ± tocopherol; 1.8 ± 1.7 years in CALM-PD L-dopa; not reported in QE2), we consider that the primary reason for the inability to detect change in most outcome measures was the presence of dopaminergic therapy. Although there was a slight increase in Motor + ADL UPDRS scores in the CALM-PD group stably treated with L-dopa over 12 months, the much larger sample size of nearly 1,200 subjects practically precludes a reasonable futility trial for subjects already treated with dopaminergic therapy using this outcome.

The 2003 Movement Disorder Society critique of the UPDRS considered the scale the "gold standard" assessment tool in PD because of its wide use, respect among physicians, and comprehensive assessment of the disorder.[17] Efforts to revise the UPDRS and improve its clinometric deficiencies currently are under–way.[17]

Time to Event Outcome Measures

A time to event outcome would also be appropriate for a futility study if a substantial percentage of control patients met a given end point over 12 months. The DATATOP trial used the onset of the need for dopaminergic therapy as the primary outcome. This outcome had a high end point rate in a short period in both DATATOP and QE2 and required only a slightly larger sample size than Total UPDRS. The need for initial dopaminergic therapy as an end point is appealing in its simplicity; however, the subjectivity of the decision to initiate dopaminergic therapy evoked debate over the outcome in the past.[18–20] Given that many features of disability unrelated to disease progression can promote the decision to initiate dopaminergic therapy, and because the sample size requirements are somewhat larger for this outcome, the time to need for dopaminergic therapy may be less suitable than the UPDRS in the futility setting. Because the futility design is an unblinded, one-arm study, the more subjective nature of this type of end point also might cause the results to be suspect.

The need for supplemental dopaminergic therapy (in patients stably treated with dopaminergic therapy) was explored as a short-term outcome measure using the CALM-PD data. This outcome had a fairly high rate of onset in 1 year, suggesting it is a feasible outcome for a short-term study. The larger sample size required for this outcome is balanced by the fact that enrollment need not be limited to untreated (not receiving dopaminergic therapy) patients. However, specific protocol demands in the CALM-PD study may have contributed to the high rate of need for supplementation. Approximately 10% of patients were unable to reach a stable dose of therapy by 10 weeks and hence were forced to receive supplemental therapy. Therefore, the true rate of need for supplemental therapy may be overestimated here. Similar to the initiation of dopaminergic therapy in untreated patients, this outcome is hampered by the inherent subjectivity of ascertainment in an unblinded study.

One must be sure of the historical rate of the onset of either initial or supplemental dopaminergic therapy, because both are likely to be influenced by changes in practice styles. If an outcome of this type were selected for a futility study, it would be prudent to include a placebo arm to verify the accuracy of the historical placebo rates in this setting and to allow blinding, even though the number of subjects would not give sufficient power to compare treatment and placebo groups.

The onset of postural instability is a clinically significant turning point in PD, one that may not be fully corrected by dopaminergic therapy. As such, it could be a viable short-term outcome measure. However, because of the frequent tendency for patients who met postural instability criteria, however measured, at one visit to revert back to normal at a subsequent visit, we found this outcome measure to be problematic for a futility study. The large number of reversions to normal over subsequent visits may be because the onset of postural instability happens gradually or because of inconsistencies in the application of the pull test, which is at the core of this assessment. A recent study evaluating the execution of the pull test indicates there is substantial variability in the method of performance, suggesting that postural instability is not well assessed.[21] When a more rigorous definition of postural instability was used in which patients must meet the definition of postural instability on at least two consecutive visits, the rates of onset were low and required too large a sample size to be pertinent for a 12-month study.

Futility Design

Although relatively new to the study of PD, futility designs have been applied to stroke and cancer research in several situations.[22,23] An underlying assumption of futility analyses is that an agent with no short-term effect will also have no long-term effect. As such, only agents that prove to be nonfutile in short-term studies will become candidates for definitive studies in long-term trials. One potential problem with this assumption is that a neuroprotective agent with a slowly developing impact on cellular death cycles may theoretically cause no observable short-term effect, may be judged as futile, and may be unfairly excluded from

further study. In this instance, if there is scientific rationale from pharmacokinetics, cellular chemistry, molecular biology, or empiric observation to doubt this caveat for a given intervention, futility studies may not be suitable. However, given the time constraints and without significant evidence that an agent is likely to have an extended development period before impact, the time frame of 6 to 18 months of follow-up is relevant for the futility setting. It is difficult to imagine neuroprotection manifesting clinically in less than 6 months, and more extended follow-up (beyond 18 months) would defeat the point of a short-term study. Given these concerns, a quickly occurring end point (1–3 months) would also not be ideal.

Another problem with focusing attention on short-term outcomes is that some elements of observed changes may reflect symptomatic effects on parkinsonism and confound the analysis of delay in disease progression. As shown in the CALM-PD study, once dopaminergic therapy is introduced, disease progression (measured by clinical rating scales) cannot be detected in small sample sizes. If a futility trial examines an agent with any symptomatic benefit, the detection of disease-modifying properties will be problematic. In this way, a purely symptomatic drug with no potential neuroprotective mechanism will prove nonfutile and remain in the candidate pool for large studies. We emphasize that futility designs do not implicitly resolve this confusion between symptomatic and neuroprotective effects. Whereas frequent visits would allow a temporal charting of changes in the UPDRS, these patterns of change would offer only insights and not clear evidence. The separation of neuroprotection from symptomatic benefit requires a trial conducted over several years.

An outcome for a futility study must convey something about disease progression at later stages to be clinically credible. If the short-term outcome used in the futility study is to be a clinical marker of disease progression, it should also correlate with long-term change. Total UPDRS as a primary outcome has been extensively used in long-term studies, showing that this measure is capable of assessing long-term change.[24–26] However, the Total UPDRS does not measure all aspects of neurodegeneration and will need to be augmented by other outcome measures for long-term trials of potential neuroprotective agents.

Limitations

Data used in these analyses were taken from available data sets from existing trials. We examined only data sets from patients with mild baseline impairment and disability. It is possible that data from other trials in different groups of patients (eg, more advanced) may give different results. Yet, patients with advanced disability would likely necessitate frequent medication changes even over 12 months, adding further analytic limitations.

Although restricting enrollment to a subset of PD patients is not ideal for neuroprotection trials, these data advise against designing short-term futility studies enrolling patients both receiving and not receiving symptomatic therapy. A weakness to enrolling only early untreated PD patients is that the impact of a neuroprotective agent in these patients may be different than for PD patients with later onset; hence, larger and longer term trials enrolling patients with more advanced disease may not see the magnitude of improvement observed in a futility study.

Given the single-arm futility design, the accuracy of the historical rate of UPDRS change observed in this study is an important consideration. Although the UPDRS is more quantitative than the decision to initiate (or supplement) symptomatic treatment, it is still susceptible to changes in practice patterns or methods of application. Furthermore, placebo rates for the revised UPDRS will need to be determined.

In this example, given the lack of established criteria, the choice of θ, the maximum allowable worsening, was relatively arbitrary (30% less than the historical control rate) and was applied universally across all outcome measures. The conclusions would be similar for any other percentage reduction, whereas a smaller percentage would require larger sample sizes. However, if there is reason to believe that the clinically meaningful difference is not a fixed percentage, but may vary across outcomes, then the outcome measure requiring the smallest sample size could change. If there is some justification for choosing different rates for different outcomes, then the approach presented in this article could be used to choose among the potential short-term outcomes using the new values for clinically meaningful differences.

LOCF was used as the method of imputation to facilitate the comparison with the results from existing studies.[14] One may wonder if the results of the test for futility would be sensitive to the method of imputation used. In the PD futility setting, LOCF may make it more likely to consider a treatment for further study compared with less biased imputation approaches, because LOCF has the potential to underestimate the magnitude of the worsening. Many other methods exist that may be less biased than LOCF, and these newer methods should be considered in designing future PD trials.[27]

Future Directions

Of those outcomes explored in this article, Motor + ADL and Total UPDRS changes at 12 months, as measured in untreated patients, are the most appropriate outcomes for a futility study. Other possible outcomes, such as the onset of the need for L-dopa, Motor

UPDRS change, and ADL UPDRS change, required sample sizes 20% greater. The onset of need for supplemental dopaminergic therapy (measured in patients stably treated with dopaminergic therapy) may also be a viable outcome if the historical rate is well estimated. If formal agreement of the definition and comprehensive training of the assessment of the pull test are established, then a short-term outcome measure for the onset of postural instability could be developed.

Futility studies of creatine, minocycline, coenzyme Q10 (CoQ_{10}), and GPI-1485 in PD patients are under way. The concept of futility studies applies to the broad range of neurological disorders for which treatment includes a focus on neuroprotection.[28] The use of short-term studies to eliminate drugs without potential allows rapid focus on agents of interest and protects the interests of patients who deserve access to the best candidates at pivotal times in the progression of their disease.

This study was supported by the grants from the NIH (National Institute of Neurological Disorders and Stroke, U01NS043127 B.C.T., and U01NS43128, K.K.).

The authors thank the National Institute of Neurological Diseases and Stroke; Steering Committees for the Deprenyl and Tocopherol Antioxidative Therapy of Parkinsonism (DATATOP),[13] Coenzyme Q10 Evaluation-2 (QE2),[14] and Comparison of the Agonist Pramipexole versus Levodopa on Motor Complications of Parkinson's Disease (CALM-PD) studies; and the Parkinson Study Group for allowing us to reanalyze their data.

References

1. Herson J. Predictive probability early termination plans for phase II clinical trials. Biometrics 1979;35:775–783.
2. Schapira A, Olanow CW. Neuroprotection in Parkinson disease: mysteries, myths, and misconceptions. JAMA 2004; 291:358–364.
3. Fahn S, Elton RL, the UPDRS Development Committee. Unified Parkinson's Disease Rating Scale. In: Fahn S, Marsden C, Calne D, et al, eds. Recent developments in Parkinson's disease. Florham Park, NJ: Macmillan Healthcare Information, 1987: 153–163.
4. Hoehn MM, Yahr MD. Parkinsonism: onset, progression, and mortality. Neurology 1967;17:427–442.
5. Schwab R, England A Jr. Projection technique for evaluating surgery in Parkinson's disease. In: Gillingham F, Donaldson I, eds. Third symposium on Parkinson's disease. Edinburgh: E. & S. Livingstone, 1969:152–157.
6. Guy W. ECDEU assessment manual for psychopharmacology-revised (DHEW pub no ADM 76-338). Rockville, MD: U.S. Department of Health, Education, and Welfare, Public Health Service, Alcohol, Drug Abuse, and Mental Health Administration, NIMH Psychopharmacology Research Branch, Division of Extramural Research Programs, 1976:218–222.
7. Ware JE, Kosinski M, Turner-Bowker DM, et al. How to score version 2 of the SF-12 Health Survey. Lincoln, RI: QualityMetric Incorporated, 2002.
8. Bushnell DM, Martin ML. Quality of life and Parkinson's disease: translation and validation of the US Parkinson's Disease Questionnaire (PDQ-39). Qual Life Res 1999;8:345–350.
9. Folstein MF, Folstein SE, McHugh PR. "Mini-mental state." A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975;12:189–198.
10. Beck AT, Ward CH, Mendelson M, et al. An inventory for measuring depression. Arch Gen Psychiatry 1961;4:561–571.
11. Hamilton M. Rating scale for depression. J Neurol Neurosurg Psychiatry 1960;23:56–61.
12. Ravina B, Fagan S, Hart R, et al. Neuroprotective agents for clinical trials in Parkinson's disease: a systematic assessment. Neurology 2003;60:1234–1240.
13. Parkinson Study Group. DATATOP: a multicenter controlled clinical trial in early Parkinson's disease. Arch Neurol 1989;46: 1052–1060.
14. Shults CW, Oakes D, Kieburtz K, et al. Effects of coenzyme Q10 in early Parkinson disease: evidence of slowing of the functional decline. Arch Neurol 2002;59:1541–1550.
15. Parkinson Study Group. A randomized controlled trial comparing pramipexole with levodopa in early Parkinson's disease: design and methods of the CALM-PD Study. Clin Neuropharmacol 2000;23:34–44.
16. Schoenfeld D. Statistical considerations for pilot studies. Int J Radiat Oncol Biol Phys 1980;6:371–374.
17. Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease. The Unified Parkinson's Disease Rating Scale (UPDRS): status and recommendations. Mov Disord 2003;18:738–750.
18. Landau WM. Clinical neuromythology IX. Pyramid sale in the bucket shop: DATATOP bottoms out. Neurology 1990;40: 1337–1339.
19. Ward C. Does selegiline delay progression of Parkinson's disease? A critical re-evaluation of the DATATOP study. J Neurol Neurosurg Psychiatry 1994;57:217–220.
20. LeWitt P, Oakes D, Cui L, et al. The need for levodopa as an end point of Parkinson's disease progression in a clinical trial of selegiline and alpha-tocopherol. Mov Disord 1997;12:183–189.
21. Munhoz R, Li J, Kurtinecz M, et al. Evaluation of the pull test technique in assessing postural instability in Parkinson's disease. Neurology 2004;62:125–127.
22. Green S, Benedetti J, Crowley J. Clinical trials in oncology. London: Chapman & Hall, 1997:44–47.
23. IMS Investigators. Combined intravenous and intra-arterial recanalization for acute ischemic stroke: the Interventional Management of Stroke (IMS) Study. Stroke 2004;35:904–911.
24. Korczyn AD, Brunt ER, Larsen JP, et al. A 3-year randomized trial of ropinirole and bromocriptine in early Parkinson's disease. The 053 Study Group. Neurology 1999;53:364–370.
25. Capildeo R. Implications of the 5-year CR FIRST trial. Sinemet CR Five-Year International Response Fluctuation Study. Neurology 1998;50:S15–S17.
26. Allain H, Destee A, Petit H, et al. Five-year follow-up of early lisuride and levodopa combination therapy versus levodopa monotherapy in de novo Parkinson's disease. The French Lisuride Study Group. Eur Neurol 2000;44:22–30.
27. Hogan J, Roy J, Korkontzelou C. Handling drop-out in longitudinal studies. Stat Med 2004;23:1455–1497.
28. Abbott A. Neurologists strike gold in drug screen effort. Nature 2002;417:109.