Assessment of Graft Effects and Function in Cell Replacement Therapy for Parkinson’s Disease

Akademisk avhandling

som med vederbörligt tillstånd
av Medicinska Fakulteten vid Lunds Universitet
för avläggande av doktorsexamen i medicinsk vetenskap
kommer att offentligen försvars i Segerfalksalen,
Wallenberg Neurocentrum, Lund,
lördagen den 4 maj 2002, kl. 09.00

av

Peter Hagell
Leg. sjuksköterska

Fakultetsopponent:
Professor Marwan I. Hariz
Neurokirurgiska Kliniken, Norrlands Universitetssjukhus, Umeå

Sektionen för Restorativ Neurologi,
Avdelningen för Neurologi,
Institutionen för Klinisk Neurovetenskap,
Lunds Universitet
This explorative study aimed to apply, develop, and evaluate assessment methods at various levels in cell replacement therapy for Parkinson’s disease (PD). Observed motor effects in five grafted patients support that timed motor tasks and clinical ratings of parkinsonism are able to monitor motor function, but their psychometric properties need further attention. Using positron emission tomography in a unilaterally grafted PD patient, dopamine receptor binding of \(^{11}\)Craclopride was found upregulated in the non-grafted, but normal in the grafted putamen. Binding reduction was normal in the grafted putamen, but not in other striatal areas, following amphetamine administration. This supports that graft-derived dopamine release can be measured in vivo in PD. The Clinical Dyskinesia Rating Scale (CDRS) yielded reliable hyperkinesia ratings, whereas dystonia ratings were reliable within, but not between, raters. In grafted patients the CDRS showed that grafts can increase dyskinesias, without associated motor improvement and unrelated to the degree of dopaminergic reinnervation. Observations support the value of the CDRS, while reliability problems call for improvements. Application of the generic health status questionnaire NHP in grafted patients demonstrated effects beyond motor function, indicating the value of such questionnaires. Psychometric evaluations of the NHP and the PD-specific PDQ-39 showed ambiguities with both, emphasizing the need for further refinements and evaluations, and illustrating the importance of systematic evaluations of outcome measures. Although most methods used here require further attention before they can be considered optimal, this study illustrates the value and importance of multi-level assessments of cell replacement therapies in PD.
Till mina föräldrar,
Majken och Roland
\alpha
**CONTENTS**

ORIGINAL PAPERS.........................................................................................................................13

ABBREVIATIONS............................................................................................................................15

SUMMARY........................................................................................................................................19

1. INTRODUCTION..........................................................................................................................23

1.1. Parkinson’s disease....................................................................................................................23

1.1.1. A brief history.........................................................................................................................23

1.1.2. Pathophysiology, pathogenesis, and etiology.........................................................................24

1.1.3. Clinical presentation...............................................................................................................24

1.1.4. Disease progression...............................................................................................................24

1.1.5. Epidemiology........................................................................................................................26

1.1.6. Pharmacological treatment....................................................................................................27

1.1.7. Complications of drug therapy.............................................................................................27

1.2. Consequences of disease and therapeutic interventions..........................................................29

1.2.1. Conceptualization..................................................................................................................29

1.2.2. Implications for assessment.....................................................................................................30

1.2.2.1. Body structure and function...............................................................................................31

1.2.2.2. Impairment (parkinsonism)................................................................................................31

1.2.2.3. Impairment (motor fluctuations and dyskinesias).............................................................32

1.2.2.4. Activity limitations and participation restrictions............................................................34

1.3. Measurement and psychometrics..............................................................................................36

1.3.1. Measurement........................................................................................................................36

1.3.2. Reliability...............................................................................................................................37

1.3.3. Validity....................................................................................................................................38

1.3.4. Responsiveness......................................................................................................................40

1.3.5. The Rasch measurement approach........................................................................................40

1.4. Cell replacement therapy..........................................................................................................42

1.4.1. Background............................................................................................................................42

1.4.2. Clinical trials..........................................................................................................................42

1.4.3. Assessment of graft survival and function............................................................................43

1.4.4. Assessment of graft effects...................................................................................................43

1.4.4.1. Parkinsonism.....................................................................................................................44

1.4.4.2. Dyskinesias........................................................................................................................44

2. OBJECTIVES..................................................................................................................................47

3. METHODS......................................................................................................................................51

3.1. Ethical approvals.........................................................................................................................51
3.2. Grafting procedures ................................................................. 51
3.2.1. Patients .............................................................................. 51
3.2.2. Tissue retrieval, procurement, and implantation ................. 52
3.2.2.1. Paper I ........................................................................... 52
3.2.2.2. Paper II .......................................................................... 53
3.2.2.3. Paper IV ......................................................................... 53
3.2.2.4. Paper V ........................................................................... 53
3.3. Assessments in grafted patients .................................................. 54
3.3.1. Motor function .................................................................... 54
3.3.2. Scanning procedures ............................................................ 55
3.3.3. Dyskinesias ....................................................................... 55
3.3.4. Perceived health and distress ............................................... 56
3.4. Development and evaluation of the Clinical Dyskinesia Rating Scale ......................................................... 57
3.4.1. Development of the scale ..................................................... 57
3.4.2. Evaluation .......................................................................... 57
3.4.2.1. The rating scale ............................................................... 57
3.4.2.2. The video tape ................................................................. 58
3.4.2.3. Raters and rating procedure ............................................. 59
3.5. Generic and disease-specific health status measurement in Parkinson’s disease ................................................................. 59
3.5.1. Design ............................................................................... 59
3.5.2. Linguistic and content validity .............................................. 60
3.5.2.1. Study sample ................................................................. 60
3.5.2.2. Evaluation ...................................................................... 60
3.5.3. Feasibility and psychometric properties ............................... 60
3.5.3.1. Study sample ................................................................. 60
3.5.3.2. Data collection ............................................................... 60
3.5.3.3. Generic health status ..................................................... 61
3.5.3.4. Disease-specific health status ....................................... 61
3.5.3.5. Additional variables ...................................................... 61
3.6. Analyses ............................................................................... 62
3.6.1. General data analytic approach .......................................... 62
3.6.2. Specific analyses ............................................................... 62
3.6.2.1. Evaluation of the dyskinesia rating scale ......................... 62
3.6.2.2. Responsiveness of outcome measures ............................ 63
3.6.2.3. Feasibility of the NHP and PDQ-39se ............................ 63
3.6.2.4. Psychometric properties of the NHP and PDQ-39se ......... 63
4. RESULTS

4.1. Motor effects of sequential bilateral DA-rich intrastriatal grafts (Paper I)
4.2. Monitoring graft function in vivo (Paper II)
4.3. Development and application of a clinical dyskinesia rating scale (papers III & IV)
4.3.1. Reliability (paper III)
4.3.2. Assessment of the influence of intrastriatal grafts of human embryonic mesencephalic tissue on dyskinesias (paper IV)
4.4. Perceived health following bilateral DA-rich intrastriatal grafts (Paper V)
4.5. Evaluation and comparison of generic and disease-specific health status questionnaires in Parkinson’s disease (Paper VI)

5. DISCUSSION

5.1. Clinical assessment of motor function
5.2. Assessment of graft function
5.3. Clinical assessment of dyskinesias
5.3.1. Development and evaluation of the Clinical Dyskinesia Rating Scale
5.3.2. Application and practical aspects of the Clinical Dyskinesia Rating Scale
5.4. Assessment of perceived health
5.4.1. Perceived health in grafted PD patients
5.4.2. Evaluation of generic and disease-specific health status questionnaires
5.5. General discussion
5.5.1. Some selected issues in assessing restorative approaches to Parkinson’s disease
5.5.2. Beyond the current paradigm
5.5.3. Implications and needs for further work
This thesis is based on the following papers, referred to in the text by their respective Roman numerals:

Copyright © Oxford University Press 1999.

Copyright © Nature America Inc. 1999.

Copyright © Movement Disorder Society 1999.


Copyright © Movement Disorder Society 2000.

VI Hagell P, Whalley D, McKenna SP, Lindvall O. Health status measurement in Parkinson’s disease: Feasibility and initial psychometric observations in a Swedish sample. *Submitted for publication.*
Paper I is reprinted by permission of Oxford University Press.
Paper II is reprinted by permission of the Nature Publishing Group.
Papers III and V are reprinted by permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADL</td>
<td>Activities of Daily Living (scale from the PDQ-39)</td>
</tr>
<tr>
<td>AIMS</td>
<td>Abnormal Involuntary Movements Scale</td>
</tr>
<tr>
<td>AMPS</td>
<td>Assessment of Motor and Process Skills</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Analysis of Variance</td>
</tr>
<tr>
<td>β-CIT</td>
<td>$[2\beta^{125}\text{T}]\text{carboxymethoxy-3}\beta(4\text{-iodophenyl})\text{tropane}$</td>
</tr>
<tr>
<td>BOD</td>
<td>Bodily discomfort (scale from the PDQ-39)</td>
</tr>
<tr>
<td>CAPIT</td>
<td>Core Assessment Program for Intracerebral Transplantations</td>
</tr>
<tr>
<td>CAPSIT-PD</td>
<td>Core Assessment Program for Surgical Interventional Therapies in Parkinson’s Disease</td>
</tr>
<tr>
<td>CDRS</td>
<td>Clinical Dyskinesia Rating Scale</td>
</tr>
<tr>
<td>COG</td>
<td>Cognitions (scale from the PDQ-39)</td>
</tr>
<tr>
<td>COM</td>
<td>Communication (scale from the PDQ-39)</td>
</tr>
<tr>
<td>CT</td>
<td>Computerized Tomography</td>
</tr>
<tr>
<td>DA</td>
<td>Dopamine</td>
</tr>
<tr>
<td>DBS</td>
<td>Deep-Brain Stimulation</td>
</tr>
<tr>
<td>DIF</td>
<td>Differential Item Functioning</td>
</tr>
<tr>
<td>DLPFC</td>
<td>Dorso-Lateral Prefrontal Cortex</td>
</tr>
<tr>
<td>DNase</td>
<td>Deoxyribo-Nuclease</td>
</tr>
<tr>
<td>EMO</td>
<td>Emotional well-being (scale from the PDQ-39)</td>
</tr>
<tr>
<td>EN</td>
<td>Energy (scale from the NHP)</td>
</tr>
<tr>
<td>ER</td>
<td>Emotional Reactions (scale from the NHP)</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>ES</td>
<td>Effect Size</td>
</tr>
<tr>
<td>FD</td>
<td>6-[^18F]-fluoro-L-Dopa</td>
</tr>
<tr>
<td>GDNF</td>
<td>Glial cell line-Derived Neurotrophic Factor</td>
</tr>
<tr>
<td>HBSS</td>
<td>Hank’s Balanced Salt Solution</td>
</tr>
<tr>
<td>HRQoL</td>
<td>Health-Related Quality of Life</td>
</tr>
<tr>
<td>HY</td>
<td>Hoehn and Yahr staging of Parkinson’s disease</td>
</tr>
<tr>
<td>ICC</td>
<td>Intra-Class Correlation</td>
</tr>
<tr>
<td>ICF</td>
<td>International Classification of Functioning, Disability and Health</td>
</tr>
<tr>
<td>ICD-10</td>
<td>10th revision of the International Classification of Diseases</td>
</tr>
<tr>
<td>ICIDH</td>
<td>International Classification of Impairments, Disabilities and Handicap</td>
</tr>
<tr>
<td>INFIT</td>
<td>Information-weighted fit statistic</td>
</tr>
<tr>
<td>IQOLA</td>
<td>International Quality of Life Assessment</td>
</tr>
<tr>
<td>IQR</td>
<td>Inter-Quartile Range</td>
</tr>
<tr>
<td>ISAPD</td>
<td>Intermediate Scale for Assessment of Parkinson’s Disease</td>
</tr>
<tr>
<td>κ</td>
<td>Kappa coefficient</td>
</tr>
<tr>
<td>K&lt;sub&gt;1&lt;/sub&gt;</td>
<td>6-[^18F]-fluoro-L-Dopa uptake rate constant</td>
</tr>
<tr>
<td>L-dopa</td>
<td>Levodopa</td>
</tr>
<tr>
<td>LFADLDS</td>
<td>Lang-Fahn Activities of Daily Living Dyskinesia Scale</td>
</tr>
<tr>
<td>Logit</td>
<td>Log-odds unit</td>
</tr>
<tr>
<td>MID</td>
<td>Minimal Important Difference</td>
</tr>
<tr>
<td>MNSQ</td>
<td>Mean-Square</td>
</tr>
</tbody>
</table>
MOB: Mobility (scale from the PDQ-39)

MOS SF-36: Medical Outcomes Study Short Form 36

MPTP: 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine

MRI: Magnetic Resonance Imaging

NHP: Nottingham Health Profile (part I)

NHPD: Nottingham Health Profile index of Distress

NMDA: N-methyl-D-aspartate

OUTFIT: Outlier-sensitive fit statistic

PA: Pain (scale from the NHP)

PD: Parkinson’s Disease

PDQ-39: The 39-item Parkinson’s Disease Questionnaire

PDQ-39se: The 39-item Parkinson’s Disease Questionnaire (Swedish version)

PET: Positron Emission Tomography

PIGD: Postural Instability and Gait Difficulty

PM: Physical Mobility (scale from the NHP)

QoL: Quality of Life

r: Pearson’s product-moment correlation

r_s: Spearman’s rank order correlation (Rho)

RAC: [¹¹C] raclopride

SD: Standard Deviation

SI: Social Isolation (scale from the NHP)
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIP</td>
<td>Sickness Impact Profile</td>
</tr>
<tr>
<td>SL</td>
<td>Sleep (scale from the NHP)</td>
</tr>
<tr>
<td>SN</td>
<td>Substantia nigra</td>
</tr>
<tr>
<td>SMA</td>
<td>Supplementary Motor Area</td>
</tr>
<tr>
<td>SOC</td>
<td>Social support (scale from the PDQ-39)</td>
</tr>
<tr>
<td>SPECT</td>
<td>Single-Photon Emission Computed Tomography</td>
</tr>
<tr>
<td>STI</td>
<td>Stigma (scale from the PDQ-39)</td>
</tr>
<tr>
<td>STN</td>
<td>Subthalamic nucleus</td>
</tr>
<tr>
<td>T</td>
<td>Kendall’s rank-order correlation coefficient</td>
</tr>
<tr>
<td>UPDRS</td>
<td>Unified Parkinson’s Disease Rating Scale</td>
</tr>
<tr>
<td>VAS</td>
<td>Visual Analogue Scale</td>
</tr>
<tr>
<td>VM</td>
<td>Ventral Mesencephalon</td>
</tr>
<tr>
<td>W</td>
<td>Kendall’s coefficient of concordance</td>
</tr>
<tr>
<td>WHO</td>
<td>World Health Organization</td>
</tr>
</tbody>
</table>
SUMMARY

Parkinson’s disease (PD) is a common neurodegenerative disorder. The core pathology of PD is progressive loss of nigrostriatal dopamine (DA) neurons leading to a striatal DA deficit, believed to cause the cardinal symptoms of bradykinesia, rigidity, tremor, and postural instability. DA-ergic drug therapy is successful early in the disease but most patients eventually develop a fluctuating drug response and dyskinesias. DA cell replacement therapy has been explored as a means to counteract PD and long-term treatment complications. Human embryonic DA-rich tissue grafted to the striatum can survive and improve motor function. The aim of this explorative study was to apply, develop, and evaluate methods for assessment of DA cell replacement therapy in PD at three different levels: body function (graft-derived synaptic DA release), impairment (parkinsonism and dyskinesias), and patients’ perceived health.

Application of timed motor tasks and clinical ratings of parkinsonism in five PD patients grafted contralaterally to a previous intrastratal transplantation revealed that the second graft gives rise to motor effects of a similar pattern but lesser magnitude than the first one. Although the assessments served to monitor motor function, their validity and psychometric properties are insufficiently documented and will require further scientific efforts.

Positron emission tomography and the DA receptor ligand \([^{11}C]\text{raclopride} (RAC) were used to monitor synaptic DA release in the presence and absence of amphetamine in a PD patient 10 years after successful unilateral transplantation to the putamen. RAC binding was upregulated in non-grafted striatal areas but normal in the grafted putamen. Similarly to healthy controls, amphetamine reduced RAC binding by 26.6% in the grafted, but only by 4.5% in the non-grafted, putamen. These observations support that graft-derived DA release can be measured in vivo in PD.

The Clinical Dyskinesia Rating Scale (CDRS) was devised to assess hyperkinesias and dystonia in various body parts. Inter- and intrarater reliability was found acceptable for hyperkinesia ratings. Dystonia ratings showed acceptable intra- but not interrater reliability. Application of the CDRS to assess graft effects on dyskinesias showed that dyskinesias can increase after transplantation, but this was not associated with improved motor function or DA-ergic reinnervation. Observed differences between hyperkinesias and dystonia, and the possibility for topographic descriptions of dyskinesias, support this approach to dyskinesia assessment, while reliability problems call for clear rating instructions and definitions, as well as for the importance of establishing rater consistency when using the CDRS.

Application of the generic health status questionnaire NHP in grafted patients revealed effects beyond those of motor symptoms, e.g., emotional reactions, energy and distress,
indicating that such questionnaires provide important information additional to that obtained by traditional clinical assessment protocols. A subsequent psychometric evaluation of the NHP and the PD-specific questionnaire PDQ-39 demonstrated ambiguities and needs for further refinements and evaluations of both questionnaires, thus illustrating the importance of systematic evaluations of outcome measures.

In conclusion, although most methods used here require further development and evaluation before they can be considered optimal, this study illustrates the value and importance of a multi-level approach to the assessment of cell replacement therapies in PD.
Too many people confuse being serious with being solemn

John Cleese
1. INTRODUCTION

1.1. Parkinson’s disease

1.1.1. A brief history

Although most certainly prevalent earlier (Stern 1989), Parkinson’s disease (PD) was first described by the London general practitioner James Parkinson in 1817 (Parkinson 1817). During the second half of the 19th century, Parkinson’s description of the disease was complemented by observations by, e.g., French neurologist Charcot, who first suggested the term PD, and his students at the Salpêtrière Hospital in Paris (Charcot and Vulpian 1861-1862; Charcot 1877, 1880, as referred in Duvoisin 1992 and Pearce 1989). Charcot (1877) also noted a mild palliative effect of a belladonna alkaloid precursor; an anticholinergic therapeutic principle that was to become the standard antiparkinsonian pharmacotherapy for the next 90 years. Somewhat later, Gowers noted a male predominance, an average age of onset between 50 and 60 years, and a familial clustering among a subset of cases (Gowers 1888, as referred in Pearce 1989). At this point, the clinical description of the disease was virtually completed. However, although the substantia nigra (SN) was suggested as the anatomic site of lesion already in the late 19th century (Blocq and Marinesco 1893; Brissaud 1895, as referred in Duvoisin 1992 and Pearce 1989), the pathological features of PD were not firmly elucidated until a couple of decades into the 20th century (Foix and Nicolesco 1925; Tretiakoff 1919, as referred in Duvoisin 1992 and Pearce 1989). While the pathophysiological consequences were still obscure, partial symptomatic relief was observed following surgical lesioning of structures within the basal ganglia as well as the thalamus during the 1940s and 1950s (Duvoisin 1992; Goetz 2001). In the late 1950s, Arvid Carlsson and colleagues identified dopamine (DA) as a neurotransmitter and showed that depletion of DA in the striatum of the basal ganglia caused a condition resembling PD that could be reversed by levodopa (L-dopa) in experimental animals (Carlsson 1959; Carlsson et al. 1957, 1958). Shortly thereafter, Ehringer and Hornykiewicz (1960) were able to verify these preclinical observations also in the human condition. These findings led to the suggestion that L-dopa could be used to treat PD (Duvoisin 1992). Whereas initial attempts to pursue this idea showed positive results (Birkmayer and Hornykiewicz 1961), subsequent replications were conflicting (Fehling 1966), and it was not until the late 1960s that L-dopa, through the work of George Cotzias and collaborators (Cotzias et al. 1967), was established as the foundation of modern PD therapy and left other therapeutic strategies, e.g., neurosurgery, largely abandoned for decades to come (Duvoisin 1992; Pearce 1989). During this time, the link between pathological findings in the SN and the DA depleted striatum was revealed through the identification of the DA-ergic nigrostriatal pathway, where nigral DA cells project to the striatum and release DA (Andén et al. 1964; Goldstein et al. 1966; Poirier and
1.1.2. Pathophysiology, pathogenesis, and etiology

Although involving a variety of anatomical sites and transmitter systems, the prime pathological alteration in PD is a loss of DA-ergic neurons in the SN leading to deficient striatal DA-ergic neurotransmission (Lang and Lozano 1998b). In untreated PD, the deficient DA innervation leads to an upregulation of DA receptors in the striatum (Seeman et al. 1987). The striatal DA paucity is believed to lead to an imbalance between the direct and the indirect striato-pallidal pathways of the basal ganglia motor circuit, resulting in clinical parkinsonism due to a net inhibitory effect of the common basal ganglia-thalamo-cortical outflow (Lang and Lozano 1998a, 1998b). The striatal DA-ergic depletion is not uniformly distributed, but less prominent in the caudate nucleus than in the putamen, and the anterior part of the putamen is less heavily denervated than the posterior part (Kish et al. 1988). These observations are also reflected in putaminal DA-ergic nerve function as measured by the uptake of 6-[18F]-fluoro-L-Dopa (FD) using positron emission tomography (PET; see below, section 1.2.2.1.) (Morrish et al. 1996a), and appear to represent differential rates of degeneration (Nurmi et al. 2001). The underlying etiology of PD is still unknown. Environmental and genetic causes are the most feasible, possibly interacting and/or comprising different subtypes of the disease (Lang and Lozano 1998a; Marsden and Olanow 1998; Williams et al. 1999). Although previously viewed as unlikely, recent studies have pointed to the involvement of genetics in the etiology of PD, at least in a subset of cases, and causative genes such as α-synuclein and parkin have been identified in a few (Kitada et al. 2000; Scott et al. 2001; Tanner et al. 1999). Oxidative stress, mitochondrial dysfunction, inflammation, deficient neurotrophic support, and excitotoxicity have been postulated mechanisms of the pathogenesis leading to cell death, and are probably, at least in part, interrelated (Lang and Lozano 1998a; Marsden and Olanow 1998).

1.1.3. Clinical presentation

Clinically, PD is characterized by four main features: paucity of movement, rigidity, tremor, and postural instability. Paucity of movement is the most characteristic motor symptom of basal ganglia dysfunction in PD (Vingerhoets et al. 1997) and is also one of the most disabling features of the disease (Lyons et al. 1998). The paucity of movement in PD refers to slowness of initiation of movement with progressive reduction in speed and amplitude of repetitive movements (Sawle 1999), thus consisting of three main entities: bradykinesia (slowness of movement), hypokinesia (reduced amplitude of movement), and akinesia (inability to initiate movement) (Marsden 1989). These terms are often used interchangeably and hereafter the term bradykinesia will mainly

Sourkes 1965, as referred in Duvoisin 1992).
be used. Many of the clinical characteristics of PD, e.g., hypomimia, micrographia, dysphagia, and respiratory difficulties, are, at least in part, various expressions of bradykinesia. Rigidity is manifested as an increase in muscle tone evident throughout the range of movement upon passive flexion and extension. Rigidity is to be considered a sign rather than a symptom; whereas it may contribute to patients’ perceived stiffness, the correlation between the two is poor (Jankovic 1992). Tremor is the symptom that most often brings the patient to seek medical help. It typically appears at rest with a 4-6 Hz frequency range and is usually most prominent in the hands but appears also in the lower extremities and the chin, whereas head tremor is rare (Sawle 1999). Although poorly correlated with the striatal DA deficit (Vingerhoets et al. 1997) and probably resulting from disruption of various components of the basal ganglia, as well as other regions (e.g., the thalamus, cortex, and cerebellum), DA neuron loss in the SN is thought to be an underlying cause (Carr 2002). Postural instability usually appears some time into the progression of disease and signifies the transition between mild to moderate PD, as defined by the Hoehn and Yahr (HY) staging of the disease (Hoehn and Yahr 1967). Although probably the least specific, postural instability appears to be one of the most disabling of the four main PD symptoms (Jankovic 1992; Lyons et al. 1998). Postural instability, as well as, e.g., disturbances of gait and speech, may, at least in part, be the result of non-DA-ergic degenerations occurring with progressive disease (Bonnet et al. 1987).

The clinical presentation of PD goes well beyond that of motor features. These non-motor features are generally less well recognized and understood. Whereas some of them to a certain extent probably are related to the primary motor disturbances, this association is less clear for others. Non-motor features of PD include, e.g., depression, neuropsychological dysfunctions, cognitive deficiencies with or without dementia, sleep disturbance, fatigue, pain and other sensory phenomena, hypotension, orthostasis, and bowel, bladder and sexual dysfunction (see, e.g., Brown et al. 1990; Cummings 1992; Edwards et al. 1992; Ford 1998; Glosser 2001; Jankovic 1992; Lou et al. 2001; Sawle 1999; Shulman et al. 2002). Furthermore, patients’ perceived PD-related problems are not solely confined to motor symptoms (Abudi et al. 1997; Brod et al. 1998; Frazier 2000). Indeed, the relationship between disease severity as judged on clinical grounds and patients’ perceived disease severity, health, and well-being have typically been found to be no more than modest to moderate (Backer 2000; Brod et al. 1998; Chrischilles et al. 2002; Fitzpatrick et al. 1997).

1.1.4. Disease progression

At the onset of PD, patients typically experience unilateral symptoms that eventually progress to bilateral involvement, although the side of symptom onset usually remains the more severely affected throughout the course of the disease (Poewe and Wenning
The clinical progression of PD can be described in terms of the HY stages (Table 1), which correlate with the number of DA-ergic cells in the SN (Ma et al. 1997) as well as with striatal DA-ergic nerve function, as measured using FD PET (Broussolle et al. 1999; Holthoff-Detto et al. 1997).

Table 1 Hoehn and Yahr staging of Parkinson’s disease (Hoehn and Yahr 1967)

<table>
<thead>
<tr>
<th>Stage</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Unilateral involvement only, usually with minimal or no functional impairment.</td>
</tr>
<tr>
<td>II</td>
<td>Bilateral or midline involvement, without impairment of balance.</td>
</tr>
<tr>
<td>III</td>
<td>First sign of impaired righting reflexes. This is evident by unsteadiness as the patient turns or is demonstrated when he is pushed from standing equilibrium with the feet together and eyes closed. Functionally the patient is somewhat restricted in his activities but may have some work potential depending upon the type of employment. Patients are physically capable of leading independent lives, and their disability is mild to moderate.</td>
</tr>
<tr>
<td>IV</td>
<td>Fully developed, severely disabling disease; the patient is still able to walk and stand unassisted but is markedly incapacitated.</td>
</tr>
<tr>
<td>V</td>
<td>Confinement to bed or wheelchair unless aided.</td>
</tr>
</tbody>
</table>

The rate of disease progression, as measured using FD PET, has been estimated to an annual decrease of about 5-7% of the normal mean FD uptake (Morrish et al. 1996b; Morrish et al. 1998). Based on these observations, Morrish and co-workers (1996b, 1998) estimated the pathological process to have begun between 3 and 7 years prior to symptom onset. This is in good agreement also with postmortem analyses of SN cell counts, yielding an estimated preclinical phase of about 5 years (Fearnley and Lees 1991). The PET-based annual progression rate largely follows estimates of the clinical progression of PD. The rate of overall clinical motor progression has thus been estimated to 4% in early PD (Poewe and Wenning 1998) and 1.5% after a mean of 6.8 years following diagnosis (Louis et al. 1999). When dividing symptomatic progression according to various symptoms, tremor did not show a clear worsening over time, whereas bradykinesia and rigidity, as well as gait and postural instability, showed an annual worsening by between 2% and 3.1% (Louis et al. 1999). These data are in agreement with clinical observations by Lee et al. (1994), which suggest a bradykinesia advance rate of 3.5% during the first year and 1.5% in the tenth year after PD onset.

1.1.5. Epidemiology

PD is a relatively common chronically progressive neurodegenerative disorder. Its age-adjusted prevalence has varied between 76 (Fall et al. 1996) and 209 (Wermuth et al. 1997) cases per 100,000 in various European studies over the past decade, with an annual incidence of about 8-15 new cases per 100,000 persons (Fall et al. 1996; Kuopio
et al. 1999), and has typically shown a slight male predominance. Mean age of onset has been estimated to between 50 and 60 years (Hoehn 1992). PD thus affects a significant minority of the population and not seldom people of working age.

1.1.6. Pharmacological treatment

Dopamine replacement by means of the DA precursor, L-dopa, is the gold standard for treatment of PD (Hassin-Baer and Giladi 2002; Lang and Lozano 1998b). L-dopa is given with a peripheral dopa-decarboxylase inhibitor (carbidopa or benserazide) to decrease peripheral conversion to DA and thereby increase bioavailability and reduce peripheral side effects. Following passage across the blood-brain barrier, L-dopa is converted to DA. L-dopa, as well as other DA-ergic drugs, is most effective in alleviating bradykinesia and rigidity, whereas its effect on tremor and postural instability is less dramatic. A number of adjunct antiparkinsonian drugs are also available. The majority of these have a DA-ergic effect, either through direct DA receptor agonism or enzyme inhibition with secondary DA- and/or L-dopa-sparing effects. The clinical efficacy of these adjunct compounds is generally inferior to that of L-dopa in long-term therapy, so that a majority of patients eventually will require L-dopa for adequate symptom control (Hassin-Baer and Giladi 2002; Lang and Lozano 1998b; Rascol et al. 2000).

1.1.7. Complications of drug therapy

Despite an often excellent initial response, most patients develop complications of therapy within some years of treatment. These can be divided into motor, autonomic, and psychiatric problems (Hassin-Baer and Giladi 2002; Lang and Lozano 1998b; Sawle 1999). The main motor complications are motor fluctuations and dyskinesias (Nutt 2001; Quinn 1998). Recent estimates based on published data indicate that both appear in about 40% of patients after 5-6 years of treatment (Ahlskog and Muerter 2001). With time, especially in patients with young-onset PD, motor fluctuations and dyskinesias often increase in severity and challenge the possibility to provide an optimal drug therapy (Quinn 1998; Quinn et al. 1987; Sawle 1999). Motor fluctuations appear as oscillations between good response to medication with good motor function and no or minimal PD-related disability (“on” phases), interrupted by episodes of poor drug response with increased PD-related disability (“off” phases) (Quinn 1998).

Dyskinesias appear as hyperkinetic and dystonic abnormal involuntary movements and postures. Dyskinesias are usually related to the “on” phase (peak-dose dyskinesias) but also appear as “off” phase or biphasic phenomena (i.e., beginning- and end-of-dose dyskinesias) (Lang and Lozano 1998b; Quinn 1998). Chorea (randomly “flowing”, purposeless, dance-like movements) is the most common type of hyperkinesias but
other types, e.g., ballism, stereotypies and myoclonus, also occur. Dystonia appear as cramp-like, brief or sustained, sometimes painful, abnormal postures, or as mobile dystonia appearing alone or mixed with chorea (Luquin et al. 1992; Marconi et al. 1994; Quinn 1998). Fixed dystonia was described long before the emergence of effective therapy (Stewart 1898) and is the only type of dyskinesia that occurs in untreated PD (Jankovic and Tintner 2001). Peak-dose dyskinesias are commonly characterized by choreiform and choreo-dystonic movements that appear spontaneously but are provoked or exaggerated by, e.g., mental stress, speaking and physical activity (Durif et al. 1999; Luquin et al. 1992; Quinn 1998; Vidailhet et al. 1999). Biphasic dyskinesias are less common and typically manifest as stereotypic, ballistic movements, and/or dystonia (Luquin et al. 1992; Quinn 1998; Vidailhet et al. 1999). In the “off” phase, sustained dystonic posturing is the predominant type of dyskinesia, although a few patients also may exhibit hyperkinesias in the “off” state (Cubo et al. 2001; Luquin et al. 1992; Quinn 1998; Vidailhet et al. 1999).

In principle, there are two prerequisites for the development of motor complications and dyskinesias in PD: repeated pulsatile treatment with DA-ergic drugs, preferably peroral L-dopa, and degeneration of the nigrostriatal DA system (Bédard et al. 1999; Schneider 1989), where the severity of the underlying disease process appears to be of greater importance than the duration of L-dopa therapy (Ahlskog and Muenter 2001; Kostić et al. 2002). Pathogenesis and pathophysiology of motor fluctuations and dyskinesias are still, however, only partly understood. Pharmacokinetic, pharmacodynamic and peripheral factors, decreased presynaptic storage capacity, dysfunctional DA release, postsynaptic alterations involving, e.g., DA and N-methyl-D-aspartate (NMDA) receptors, and opiate transmission, altered striatal expression of Fos B-related proteins and neuropeptides, may all contribute (Bédard et al. 1999; Bezard et al. 2001).

The need for novel therapeutic interventions that are able to offer relief where traditional therapy fails to do so is thus apparent. In this respect, new promising DA-ergic and non-DA-ergic compounds, as well as neurosurgical interventions, have rendered great interest lately (Colzi et al. 1998; Ferreira and Rascol 2000; Hassin-Baer and Giladi 2002; Lang and Lozano 1998b). Whereas all currently established therapeutic interventions for PD are symptomatic, by compensating either for the deficit in striatal DA transmission or for the secondary imbalance downstream in the basal ganglia, none has been proven to slow disease progression or restore the degenerated nigrostriatal system.
1.2. Consequences of disease and therapeutic interventions

1.2.1. Conceptualization

The International Classification of Impairments, Disabilities and Handicap (ICIDH), first published by the World Health Organization (WHO) in 1980 (WHO 1980), later revised (WHO 1997) and recently appearing as the International Classification of Functioning, Disability and Health (ICF) (WHO 2001), together with the 10th revision of the International Classification of Diseases (ICD-10) (WHO 1992-1994), constitutes a practical classification of disease, health, and consequences of ill health. The scope of the ICF has been broadened, as compared to the ICIDH, from defining only “consequences of disease” to encompass also “components of health”. The ICIDH thus defined the consequences of “health conditions” (diseases) as “impairment” (loss or abnormality of body structure or function, broadly conceptualizing symptoms and signs), “disability” (restriction or lack of performing activities due to an impairment), and “handicap” (disadvantages that limits or prevents a person to fulfill his/her normal role due to an impairment or disability). In the ICF, each level has been expanded to comprise also a positive counterpart, referred to as “functioning”. The nomenclature of the various levels of consequences of disease has also been partly revised to comprise “impairment”, “activity limitation”, and “participation restriction”, which are defined as physiological or anatomical problems in body functions or structure, difficulties in executing activities, and problems in involvement in life situations, respectively, where the latter two are considered as two facets of the same level and disability serves as an umbrella term for all three. In addition, it is recognized that the presence or degree of disability is influenced not only by the underlying health condition, but also by various environmental factors (facilitators/hindrances).

In terms of PD (Table 2), the ICF/ICD-10 framework can be operationalized to comprise (a) deficient striatal DA neurotransmission (loss of body function) due to degeneration of the nigrostriatal system (loss of body structure), giving rise to (b.i.) clinical symptoms and signs (impairments) constituting the motor and non-motor features of PD, and, in combination with DA-ergic drug therapy, motor fluctuations and dyskinesias. These impairments may, in interaction with various environmental factors (e.g., social, physical and psychological), impose various consequences for the affected person’s daily life (b.ii.) in terms of altered abilities to perform activities and participate in his/her normal life situations (e.g., self-care, house keeping, work, and leisure activities).

Because disease may have consequences on all levels outlined above, it can also be presumed that interventions aimed at counteracting disease or alleviating its consequences also may have the potential to influence various facets thereof. For example, if a certain symptom is the chief cause of patients’ limited abilities to function
on a day-to-day basis and remain in employment, then it would be reasonable to assume that effective treatment of that symptom will be capable not only of producing symptomatic relief, but also of improving reduced activity limitations and participation restrictions. Depending on the type and mechanism(s) of the intervention, it may or may not also influence the underlying change in body structure or function. However, the relationship among various consequences of disease is not linear, and one level does not necessarily strictly follow or depend on the previous one (McKenna et al. 2000a). Indeed, this was one of the major criticisms of the first version of the ICIDH framework (WHO 1980). For example, studies in stroke have indicated that functional improvements following rehabilitation only in part can be explained by altered neurologic symptomatology (Roth et al. 1998). Similar observations have been made in PD, with patients experiencing functional deteriorations in the absence of altered clinical measures of parkinsonism (Fitzpatrick et al. 1997).

1.2.2. Implications for assessment

Because all levels outlined in section 1.2.1. may be influenced by disease as well as treatment, they also bear implications when attempting to assess and measure the effects of therapeutic interventions (Hobart et al. 1996a). Furthermore, since the relationships between various levels are not linear, assessment of one level does not yield valid information on the other. Measures thus need to focus on one facet at a time and various approaches need to be used depending on the studied construct. This is one of the basic assumptions underlying valid and objective measurement (see also below, sections 1.3. and 5.5.1.). Assessment of disability (i.e., impairment, activities, and participation) in PD poses additional aspects of parkinsonism, motor fluctuations, and dyskinesias that need to be taken into consideration: (a) the amount present at a particular time, i.e., severity in terms of impairment; (b) parkinsonian symptomatic profile and types of dyskinesias present (e.g., hyperkinetic vs dystonic dyskinesias); (c) daily

<table>
<thead>
<tr>
<th>Conceptual level</th>
<th>Parkinson’s disease</th>
<th>Mode(s) of assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Body structure and function</td>
<td>Nigral DA cell degeneration with striatal DA deficiency</td>
<td>In vivo brain imaging</td>
</tr>
<tr>
<td>(b) Disability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(i) Impairment</td>
<td>Bradykinesia, rigidity, tremor, postural instability Motor fluctuations, dyskinesias Non-motor symptoms</td>
<td>Objective quantification Clinical examination Self-assessment</td>
</tr>
<tr>
<td>(ii) Daily living</td>
<td>Activities: Altered abilities to perform activities (e.g., P-/I-ADL) Participation: Altered possibilities for participation in life situations</td>
<td>Clinical observation Self-assessment</td>
</tr>
</tbody>
</table>

Table 2: Consequences of disease and therapeutic effects in Parkinson’s disease

a Common term for activities and participation not included in the International Classification of Functioning, Disability and Health (WHO 2001) but used here for practical reasons.
DA, dopamine; P-ADL, personal activities of daily living; I-ADL, instrumental activities of daily living
duration (of, e.g., “off” phases and dyskinesias); and, (d) influences on daily living (i.e., activity limitation and participation restriction). Various principal modes of assessment with relevance to PD at each level are outlined in Table 2 and specific approaches are briefly reviewed below.

1.2.2.1. Body structure and function
The core structural pathology of PD, i.e., loss of striatal DA innervation due to nigral DA cell degeneration, can indirectly be monitored in vivo using either of two imaging techniques, single-photon emission computed tomography (SPECT) or PET, typically using the presynaptic ligands $[^2\beta\text{CIT}]$ and FD, respectively. Both SPECT and PET have been shown to yield comparable and reliable estimates of striatal DA deficiency correlating with the severity of clinical PD symptoms (Asenbaum et al. 1996; Ishikawa et al. 1996; Morrish et al. 1996b; Seibyl et al. 1997; Vingerhoets et al. 1994), although SPECT has a poorer resolution (Marek 1999).

Using PET, FD is decarboxylated and its presynaptic uptake is measured and expressed as the uptake rate constant ($K_i$), thus providing a measure of the number of viable striatal DA terminals (Brooks 2000). The reliability (ICC, intra-class correlation) of striatal $K_i$ has been estimated at between 0.80 and 0.91 for various striatal structures (Vingerhoets et al. 1994). In PD, putaminal FD uptake correlates inversely with the degree of motor impairment (Morrish et al. 1996b; Remy et al. 1995), whereas the correlation between motor symptoms and decreases in FD uptake in the caudate nucleus is weaker (Holthoff-Detto et al. 1997; Morrish et al. 1996b). Striatal FD uptake in vivo correlates strongly with post-mortem findings regarding numbers of remaining substantia nigra DA neurons as well as striatal DA levels (Snow et al. 1993), providing sound basis for the validity of the method. Using PET, it has also been possible to monitor the postsynaptic striatal upregulation of DA receptors in drug-naive PD in vivo by measuring striatal binding of the DA D2 receptor ligand $[^{11}\text{C}]$raclopride (RAC), which also has revealed that the receptor upregulation is reversed following chronic L-dopa treatment (Antonini et al. 1994).

The functional pathology of the disease process in PD, i.e., the deficient striatal DA-ergic neurotransmission, is not reflected by FD PET. However, because RAC is a low-affinity, reversible D2 receptor antagonist that is displaced following increased synaptic DA concentrations, it has been possible to quantify changes in synaptic DA concentrations induced by psychostimulant drugs or arousal in non-parkinsonian subjects (Breier et al. 1997; Koepp et al. 1998; Volkow et al. 1994), as well as following L-dopa administration in PD (Tedroff et al. 1996).

1.2.2.2. Impairment (parkinsonism)
Assessment and measurement of the symptoms and signs of PD are typically based on clinical examination. Numerous clinical rating scales have been suggested for quantification and assessment of symptomatology and efficacy of therapeutic interventions (for review, see Marínez-Martín 1993). Of these, the Unified Parkinson’s Disease Rating Scale (UPDRS) (Fahn et al. 1987) has become the golden standard (Mitchell et al. 2000). The UPDRS consists of 42 items divided into four scales: Mentation, Behavior, and Mood (section I; 4 items); Activities of Daily Living (section II; 13 items); Motor Examination (section III; 14 items); and, Complications of Therapy (section IV; 11 items). In addition, the Schwab and England Activities of Daily Living scale and a modified HY staging are included. Section III, the motor examination, is the most widely used and evaluated clinical rating scale of overall parkinsonian motor impairments. It covers various aspects of the main symptoms of PD, thus allowing for basic symptomatic profiling as well as severity assessment, and experiences from the past 14 years support its general usefulness and validity (Fahn et al. 1987; Goetz et al. 1995; Lang 1995; Martínez-Martín et al. 1994; Morrish et al. 1996b; Richards et al. 1994; Seibyl et al. 1997; Siderowf et al. 2002; Stebbins et al. 1998, 1999).

A general concern when using clinical rating scales is their subjectivity; that is, the score is dependent not only on the motor performance of the patient, but also on the examiner’s interpretation thereof. More objective quantification methods, such as the Postural-Locomotor-Manual test (Steg et al. 1989), have been developed. Such technical devices are, however, relatively expensive and their use may require special location, equipment, and training, which limits their usefulness as routine clinical assessment tools. As a compromise, various timed motor tests that are easy to perform and do not require any technical equipment beyond a stopwatch, have been described (Hagell 2000; Langston et al. 1992; Morris et al. 1998a, 1998b). The tapping test, or hand/arm movement between two points, is the timed upper extremity test that has been most widely applied and best evaluated. When performing this test, the patient is required to successively tap two points separated by a standardized distance in front of him/her and the number of taps performed within a predefined time frame is recorded (alternatively, the time taken to perform a predefined number of taps can be recorded). This test has been demonstrated to be a highly sensitive and specific measure of bradykinesia and DA-ergic responsiveness (Boraud et al. 1997; van Hilten et al. 1997). Timed walking tests over a defined distance are also commonly used and have been shown to be highly reliable and able to differentiate between PD patients and age-matched healthy controls (Morris et al. 2001).

1.2.2.3. Impairment (motor fluctuations and dyskinesias)
Other important consequences of PD and its therapy that need consideration in terms of assessment are motor fluctuations and dyskinesias. Because dyskinesias take a variety of clinical expressions (see above, section 1.1.7.), these have to be considered also when attempting to assess them (see above, section 1.2.2.). Several scales have been
suggested and used to assess various aspects of dyskinesias in PD and some of the more common approaches will be considered below (for more detailed reviews, see Goetz 1999; Hoff et al. 1999).

The amount of dyskinesias present, i.e., its severity in terms of impairment, has mainly been assessed using the abnormal involuntary movements scale (AIMS) (Guy 1976, as referred in Marsden and Schachter 1981) or various modifications thereof (e.g., Davis et al. 1991; Limousin et al. 1995; Uitti et al. 1997). AIMS evaluates the severity of observed dyskinesias on a non-defined scale ranging from 0 (= none) to 4 (= severe) in each of seven body parts: muscles of facial expression, lips and perioral area, jaw, tongue, upper extremities, lower extremities, and neck, shoulders and hips. The scale was not devised for PD, but to assess tardive dyskinesias, which explains its bias toward the orofacial area. While evaluated for its intended use, i.e., to assess tardive dyskinesias (e.g., Barnes and Trauer 1982), AIMS has not been evaluated for use in PD, although occasional investigators have provided rater reliability when using modifications of the scale (Durif et al. 1997). AIMS does not address what type(s) of dyskinesias the patient exhibits. Another approach to the assessment of the amount of dyskinesias present is the use of various ambulatory technical devices. However, besides the practical limitations generally associated with such techniques (see above, section 1.2.2.2.), an additional problem in relation to dyskinesias is the ability of devices to distinguish between normal motor activity, tremor, and dyskinesias (Brown and Manson 1999; Hoff et al. 2001).

There are different approaches to obtain information on the daily duration of dyskinesias and motor fluctuations. One commonly employed method is to instruct the patient to keep “on”/“off” diaries differentiating between, e.g., “off”, “partial on”, “on”, and “on with dyskinesias” at regular intervals throughout the day (Langston et al. 1992), yielding information on the amount of time spent in the various conditions. “On”/“off” diaries are a highly subjective means of assessment, but good agreements between patients’ and clinicians’ assessment can be achieved by training (Goetz et al. 1997; Parkinson Study Group 2001). An alternative approach is for the clinician to estimate the time spent in “off” and with dyskinesias as based on retrospectively obtained patient, family and/or caregiver history. This approach is employed in the fourth section of the UPDRS (UPDRS IV, complications of therapy), where the daily amounts of time spent in “off” and with dyskinesias are classified into either of five categories (0%; 1-25%, 26-50%; 51-75%, or 76-100%) (Fahn et al. 1987). The main problem with this approach is the fact that it largely relies on recall (Streiner and Norman 1995) and that categories are relatively broad.

The Obeso dyskinesia rating scale (Langston et al. 1992) consists of a mixture of the daily duration of dyskinesias and their influence on daily activities. The scale has two parts, one concerns dyskinesia intensity (in terms of activity limitation, score range: 0-
5) and the other daily duration in a manner similar to that of the UPDRS (score range: 0-5), with the arithmetic mean of the two as the scale score. This scale appears to lack formal evaluation, has rarely been used, and has received severe criticism (Goetz 1999).

1.2.2.4. Activity limitations and participation restrictions

Consequences of disease related to daily living can be assessed either by clinical observation or self-assessments. As opposed to symptom severity (Golbe and Pae 1988; Hagell and Sandlund 2000), patients’ self-assessments of functional limitations are generally in accordance with that provided by observers (Brown et al. 1989). A variety of observer-derived measures relating to various aspects of daily living are available (McDowell and Newell 1996). A few of these have been tailored specifically towards PD. One of the more well known PD specific instruments is the Northwestern University Disability Scale (Canter et al. 1961), which was one of the more frequently used scales up until about ten years ago. Since then, the activities of daily living section of the UPDRS (UPDRS II) has become the most commonly used scale. However, the validity of UPDRS II has been seriously questioned, primarily due to the fact that it encompasses a mixture of impairments and activity limitations (van Hilten et al. 1994). A more recent and promising development in terms of PD specific clinical rating scales of activity limitations is the Intermediate Scale for Assessment of PD (ISAPD) (Martínez-Martín 1993; Martínez-Martín et al. 1995). One generic observer-derived objective measure of instrumental activities of daily living is the Assessment of Motor and Process Skills (AMPS), which recently has been applied in PD with encouraging results (Corbett 1998; Hariz et al. 1998). AMPS probably represents the best and most objective available measure of performance of daily activities (Josman and Birnboim 2001), but its more general use is somewhat hampered by the need for specially trained occupational therapists.

Specifically addressing the contribution of dyskinesias to PD patients’ activity limitations, Goetz and collaborators (1994) suggested and evaluated the Rush dyskinesia scale as a means to assess the severity of dyskinesias based on interference in activities of daily living (Goetz et al. 1994). Assessments are prescribed to be based on observation of the patient performing three standardized activities (walking, drinking from a cup, and putting on and buttoning a coat). One global, non-lateralized, score ranging from 0 (no dyskinesias) to 4 (violent dyskinesias, incompatible with any normal motor task) is used. Inter- and intrarater agreement of these ratings reached generally acceptable levels (Kendall’s $W = 0.710 - 0.897$ and Spearman’s Rho $[r_s] = 0.826 - 0.908$, respectively). In addition, the rater should distinguish the types of dyskinesias present, classified as chorea, dystonia and “other”, and identify which one of these that is the most disabling. Whereas intrarater agreement for identifying the latter was acceptable (Cramér’s $V = 0.836 - 0.840$), it was suboptimal for classification of the types of dyskinesias present (Cramér’s $V = 0.699 - 0.732$). Interrater agreement for both classification and identification of the most disabling type of dyskinesias were low.
(Kappa coefficient [$\kappa$] = 0.279 - 0.483). More recently, another scale for specific assessment of dyskinesia-related activity limitations, the Lang-Fahn Activities of Daily Living Dyskinesia Scale (LFADLDS), was presented (Parkinson Study Group 2001). This scale is a modification of five of the items in the activities of daily living section of the UPDRS (Fahn et al. 1987). The patient is required to respond, using a 5-grade scale (0, no dyskinesias; 4, incapacitating dyskinesias), to questions attempting to assess how dyskinesias, at maximum severity, influence his/her ability to write, eat, dress, attend to hygiene, and walk. In comparison with other dyskinesia indices, the LFADLDS has correlated weakly to moderately with the time spent in “on” with dyskinesias, as recorded by patients at home ($r_s = 0.33 - 0.46$) and by patients and investigators during clinical observation ($r_s = -0.02 - 0.56$), indicating that the two tap different facets. The correlation between the LFADLDS and the Rush dyskinesia scale, also aiming to quantify dyskinesia-induced activity limitations but as assessed by an examiner, was, however, absent to weak ($r_s = 0.02 - 0.36$), thus questioning the validity of either of the scales. No information regarding the reliability of the LFADLDS appear to yet have been published. Finally, the influence of dyskinesias on daily activities is also recorded by one item of the UPDRS IV, which classifies dyskinesias-induced disability as none, mild, moderate, severe, or complete (Fahn et al. 1987).

There is an increased need and demand for outcome measures that capture patients’ perspective (Devinsky 1995; Hobat et al. 1996a). The most widely employed approach to patient-reported outcomes measurement thus far is the use of health status or health-related quality of life (HRQoL) questionnaires. Patrick and Erickson (1993) have defined HRQoL as “…the impairments, functional states, perceptions and social opportunities that are influenced by disease, injury, treatment or policy”, thus broadly operationalizing the WHO’s definition of health (WHO 1958) as “…physical, mental, and social well-being, and not merely the absence of disease or infirmity”. The approach thus broadly operationalizes disability, as defined by the ICF (WHO 2001). The term HRQoL has over the past decade come to be used for the same measures that previously were known as health status questionnaires and is regarded equal to health status and functional status (Bergner 1989; Guyatt et al. 1993); they are often used interchangeably with one another as well as with quality of life (QoL). However, because accumulating evidence point to the invalidity of equating health and QoL (Bowling and Windsor 2001; Engström and Nordeson 1995; Michalos et al. 2000, 2001; Mozes et al. 1999; Smith et al. 1999; Stensman 1985), the term health status, or perceived health, is preferred and will be used here. Although incorporating also impairments, health status questionnaires largely focus on the levels of performance of various activities and participation within various areas of life (McKenna et al. 2000a). They are typically multidimensional, consisting of different scales that tap various physical, mental, and social aspects of health. Health status questionnaires can be classified as generic or disease-specific (Patrick and Deyo 1989). Generic questionnaires are broad in scope and have been developed for use among a wide range of patient populations, as well as
in healthy individuals. Therefore, such instruments may lack in relevance, sensitivity, and responsiveness among patients with a particular condition (Patrick and Deyo 1989; Fayers and Machin 2000). This has led to the development of disease-specific questionnaires. Several PD-specific health status questionnaires have been developed, among which the 39-item PD Questionnaire (PDQ-39) (Peto et al. 1995) is the most widely used to date (Marinus et al. 2002). Among generic questionnaires, the Nottingham Health Profile (NHP) (Hunt et al. 1980; Hunt and McKenna 1989), the Sickness Impact Profile (SIP) (Bergner et al. 1981), and the Medical Outcomes Study Short Form 36 (MOS SF-36) (Ware and Sherbourne 1992) are the more commonly used in PD (Peto and Jenkinson 1999). In addition to psychosocial aspects of activity limitations and participation restrictions, these questionnaires generally encompass much of the same, or very similar, aspects that are covered by the many observer-derived or self-assessed scales previously devised as measures of activities of daily living (Canter et al. 1961; Everitt et al. 1989; Martínez-Martín 1993), among which observer-derived assessments generally have not displayed any apparent advantages over self-assessments (Brown et al. 1989; Geminiani et al. 1991; Henderson et al. 1991).

1.3. Measurement and psychometrics

1.3.1. Measurement

The history of measurement is long and its development can be traced through a variety of sciences such as physics, mathematics, and psychology, typically generating fundamentally common rules and principles (Bond and Fox 2001; Wright 1997; Wright 1999). The traditional broad definition of measurement has been “the assignment of numerals to objects or events according to rules” (Stevens 1946). As an extension of this definition, Stevens (1946) outlined four different kinds of measurement scales according to the rules of assigning numerals, the mathematical properties of, and statistical operations applicable to the resulting scales. While criticized and subjected to variations and extensions (see, e.g., Stine 1989; Wright 1999), Stevens’ classification of measurement scales has served as a general standard for the past half-century. The nominal scale thus consists of mere labeling, where the numerical labels bear no quantitative meaning, whereas the ordinal scale arises from rank ordering where the assigned numbers (e.g., 0 - 1 - 2 - 3) tell us that “1” is more than “0” and that “2” is more than “1” but less than “3”, but contains no information about the distances or magnitudes between the numerals used. The two final scale types, the interval and ratio scales, also give this information but in addition they are linear, i.e., their intervals are equal so that the distance between “0” and “1” is the same as that between “2” and “3”. The distinction between the interval and the ratio scale lies in that whereas the former does not have an absolute zero, this is implied in the latter.
Attempts to measurement in clinical medicine usually involve consequences of ill health and bodily dysfunction, and the effects of different interventions thereon. In order to measure something in a clinically and scientifically defensible manner, any measurement tool, whether used to measure symptom severity, cerebral uptake of radioactively labeled ligands or perceived health, must fulfill certain criteria in order to be considered useful and appropriate for clinical and scientific conclusions and decisions. In addition to clinical appropriateness, certain minimum standards regarding the psychometric (i.e., scientific) properties of measurement tools must thus be documented and met (Hobart et al. 1996b; McDowell and Newell 1996; Streiner and Norman 1995). Psychometrics has its roots in the social sciences, particularly in psychology, where it has undergone an impressive development during the past century but only relatively recently been incorporated into the area of clinical medicine (Hobart et al. 1996b; McDowell and Newell 1996). The essential psychometric properties in clinical research can broadly be classified into reliability, validity, and responsiveness.

1.3.2. Reliability

Reliability refers to the precision of a measure, i.e., the extent to which it is free from measurement error. Among several different types, test-retest reliability and internal consistency are some of the most common. For observer-derived measures it is also important to consider interrater reliability (i.e., to what extent results are reproduced among observers rating the same situation) and intrarater reliability (i.e., to what extent results are reproduced when one observer rates the same situation at different time points). Test-retest reliability can be seen as an equivalent to intrarater reliability for self-administered measures.

Test-retest reliability estimates to what extent results are reproducible over time among subjects who have not changed on the measured domain(s). This requires that the measured construct is stable between the two administrations. Therefore, to avoid the influence of actual change, the time between two administrations should not be too long. Conversely, if the time lag is too short, the results from the second administration is likely to be influenced by respondents’ (or raters’) recall from the first occasion. As a compromise, a general recommendation is that one to two weeks should separate the two test occasions (Bowling 1997a; Deyo et al. 1991; Nunnally and Bernstein 1994; Streiner and Norman 1995). Test-retest reliability is commonly estimated by means of correlation analyses according to Spearman ($r_s$) or Pearson ($r$), where a coefficient of 1.0 is interpreted as perfect stability and 0 as no stability over time or between raters. A problem with such methods is that they estimate linear co-variation and not agreement. Hence, a systematic difference of a certain magnitude between the first and second occasion is not reflected by these statistics. Although this can be assessed by use of
post hoc comparisons, it is often recommended to use intra-class correlation (ICC) instead (Deyo et al. 1991; Fleiss 1986; Streiner and Norman 1995). ICC is based on repeated measures analysis of variance (ANOVA) and incorporates both co-variation and agreement into its calculation. Test-retest reliability is essential to judge the usefulness of an instrument and some sort of information regarding this is necessary for any decision regarding the value of a particular measurement tool (Fayers and Machin 2000; Streiner and Norman 1995).

Internal consistency refers to the reliability of cross-sectional data and is commonly estimated using Cronbach’s coefficient alpha (α) (Cronbach 1951), which can take any value between 0 and 1, where 1 indicates maximal consistency. Coefficient α is an extension of the older “split-half” technique, but instead of splitting the items into two halves, α yields an estimate of the average of all possible split-half combinations (Nunnally and Bernstein 1994). In contrast to what often is believed, coefficient α is not a measure of unidimensionality, but of item interrelatedness and an estimate of reliability (Clark and Watson 1998; Cortina 1993; Fayers and Machin 2000). Furthermore, coefficient α is dependent not only on items’ average inter-correlation, but also on the number of items (Cortina 1993). Increasing the number of items can thus be one way of increasing internal consistency, as well as other forms of reliability (Nunnally and Bernstein 1994; Streiner and Norman 1995).

Reliability coefficients should be at least 0.80 for an instrument to be used at the group level, whereas the standard at the individual level is more stringent and should be at least 0.90 (Clark and Watson 1998; Fayers and Machin 2000; Nunnally and Bernstein 1994; Streiner and Norman 1995). These standards have been derived to enable valid inferences based on assessment instruments and to minimize the potentially serious consequences of unreliability in the design and conduct of clinical experimental and non-experimental research. These consequences are discussed and illustrated in some detail by Fleiss (1986) and include, e.g., attenuated correlations between measures, demands for larger sample sizes, and biased sample selection.

1.3.3. Validity

Whereas reliability estimates an instrument’s precision, it does not tell us what it measures. To address this, its validity must be evaluated. The approaches to evaluation of validity are plenty and depend on the type of measure or assessment method that needs validation. In assessment of observable variables, validity is primarily dependent on clear consensus or definitions regarding the attributes or expressions of the variable. Establishing validity for measures of latent (i.e., non-observable) variables is often more intricate. Although generally applicable, the following section will primarily refer to latent trait validity. Numerous types of validity have been described over the
years, but the most common ones are face, content, construct, and criterion validity (Nunnally and Bernstein 1994; Streiner and Norman 1995). Face validity relates to how the assessment tool and its items appear in relation to the intended use and construct, whereas content validity is a more systematic evaluation of its content and items. Content validity is often evaluated through critical review of the assessment tool’s comprehensiveness by a panel of experts and/or patients, depending on the assessed level and construct. Construct validity is evaluated by a priori hypotheses formulations regarding how a measure should “behave” in relation to other indices and/or among subgroups of subjects. These hypotheses are subsequently evaluated by, e.g., correlation analyses between measures and/or comparison of results among subsets of subjects. Construct validation should be viewed as a continually ongoing process where empirical results build on evidence and feed revisions and further testing; it can never be fully established in a single study. In criterion validation, the new instrument is tested against an existing and established criterion or “gold standard” measure of the same construct. Criterion validity is typically addressed when a shorter instrument is developed to replace an existing but longer one.

Many validity estimates are based on correlation coefficients. Therefore, the interpretation of such coefficients needs to be considered. Statistical significance levels are often used as a basis for interpretations and conclusions. This approach should, however, be viewed with caution because the tested null hypothesis is that the studied measures do not correlate, i.e., bear a correlation coefficient that does not differ from zero. Furthermore, the sample size contributes profoundly to whether the null hypothesis will be rejected. Hence, any correlation coefficient above or below 0.0 can yield statistical significance, given that a large enough sample is studied (Norman and Streiner 2000). Therefore, it is important to consider the strength of the actual coefficients instead of relying on p-values when interpreting results. The strength of correlation coefficients <0.20 can thus be considered poor; 0.20-0.35 slight; 0.35-0.65, moderate; 0.65-0.85 good; and >0.85 very good (cf. Cohen and Manion 1994).

In order to fully enable establishment of validity, any instrument should be unidimensional, i.e., it should tap only one distinct underlying construct. Furthermore, this construct should be clearly defined and based on a stated theory or model. Otherwise, appropriate pre-testing hypotheses, as well as valid interpretations of, and conclusions based on, test results cannot be made (Bond and Fox 2001; McDowell and Newell 1996; Mishel 1998). When several constructs need to be considered, they should be so by various unidimensional approaches (e.g., various subscales within a multidimensional health status questionnaire) rather than being confounded within a single index (Nunnally and Bernstein 1994). An objective approach to establishment of unidimensionality, and, indeed, of most psychometric properties in latent trait measurement, is the use of Rasch analysis (see below, section 1.3.5.).
1.3.4. Responsiveness

Outcomes measurement in clinical health care is often concerned with changes over time, in particular as a result from therapeutic interventions. Therefore, responsiveness, i.e., the ability to detect small but clinically important changes, is a paramount feature of instruments used for this purpose (Deyo et al. 1991; Hobart et al. 1996b). Various estimates of responsiveness have been suggested. These can be classified into either of two broad strategies, anchor- or distribution-based approaches. Examples include the anchor-based minimal important difference (MID) and the distribution-based effect sizes (ESs) (Deyo et al. 1991; Fayers and Machin 2000; Husted et al. 2000; Kazis et al. 1989; Norman et al. 2001). The MID represents the smallest change in score that patients perceive as beneficial and/or that causes clinicians to consider a change in patient management. ESs are based on the distribution and variability of data. One general advantage is that ESs allow for comparisons across various measures, despite different absolute score values and ranges (Fayers and Machin 2000). Several variations of ESs have been suggested, among which the standardized ES (referred to as ES hereafter) is one of the more common. The ES is calculated as the mean score change divided by the baseline standard deviation (SD). Norman and collaborators recently demonstrated that the ES is largely independent on the cut-off score chosen for a MID and is able to predict a likely proportion benefiting from an intervention, whereas the relation between MID and benefit shows anomalies, indicating that using the MID in interpretation of change at the group level is problematic (Norman et al. 2001). As a suggested rule of thumb, ESs of 0.20 to 0.49 are regarded small; 0.50 to 0.79 moderate; and 0.80 or above, large (Fayers and Machin 2000; Kazis et al. 1989).

1.3.5. The Rasch measurement approach

Rasch analysis (Bond and Fox 2001; Rasch 1980; Smith 2001; Wright 1997; Wright and Masters 1982; Wright and Mok 2000; Wright and Stone 1979) allows assessment of fundamental measurement characteristics, such as unidimensionality (i.e., whether a test measures one underlying construct), relative item difficulties (i.e., the distance in interval measurement units between items), and hierarchical item order (i.e., item ordering along the underlying unidimensional construct). This section will briefly outline the basic principles of the Rasch measurement approach, and exemplify some of the analytic approaches it offers.

According to the Rasch model, the probability of a person giving a certain response to an item is a logistic function of the difference between item “difficulty” and respondent “ability”. In terms of, e.g., health, this translates to the level of health defined by the item and the respondent’s health status. The hierarchical order of items is defined by
item calibrations expressed as “logits” (log-odds units) along the hierarchical scale. The location of a particular item represents its severity relative to that of the total set of items. The logit metric ranges from minus infinity to plus infinity, with item mean difficulty set at zero. The distance between two logits is the distance along the line of inquiry that increases the probability of observing the specified event (e.g., the problem defined by an item) by a factor of 2.718. This distance is the same throughout the continuum. The logit metric is thus linear and performs at the interval level (Stevens 1946).

Persons are measured along the same logit metric as items, which, e.g., allows for direct comparison of how well a certain test captures the range of performance among its examinees. Because items that are endorsed by all or none of the examinees, and examinees that endorse all or none of the items, do not provide any information regarding the true item and person measures (except of being out of range), such observations are omitted (although rough estimates can be provided). Hence, the Rasch model does not only provide interval measures for items and persons but yields information freed from the influence of the particular set of people and items used, and is, therefore, a fundamental model for objective measurement. If items are appropriately targeted for the investigated sample, the mean ability measure of the sample should approximate the mean item difficulty measure (i.e., around 0 logits).

Unidimensionality can be assessed by determining each item’s goodness-of-fit, i.e., to what degree each item of a scale contributes to the measurement of the same underlying construct. There are two different fit statistics for the Rasch model, the information-weighted (INFIT) and the outlier-sensitive (OUTFIT), which measure fit near and further away from the level of observed responses, respectively. These can be expressed as mean-square (MNSQ) and standardized Z statistics, of which the Z statistic is more robust across various sample sizes and numbers of items than the MNSQ statistic (Smith 2000; Smith et al. 1998).

Rasch analysis also provides item and person separation indices, defined as the ratio of the true spread of the measures with their measurement errors. These can be used to determine if a measure yields sufficient spread along the underlying continuum to define distinct levels (strata) of item difficulties and person abilities (Smith 2001; Wright and Masters 1982).

Furthermore, because the Rasch models of measurement are “sample free”, i.e., item difficulty estimates are independent of the sample, they provide valid means of evaluating differential item functioning (DIF) (Leplége and Ecosse 2000; Scheuneman and Subhiyah 1998; Smith 1992). Assessment of potential DIF is important because if a measure behaves different across various subsets of examinees, e.g., gender and age
groups, this will challenge the validity of its use when pooling and comparing data from such subgroups (Bond and Fox 2001; Fayers and Machin 2000; Whalley 1996).

1.4. Cell replacement therapy

1.4.1. Background

The idea to restore function and repair the brain by grafting neurons is not new. The first attempts to do so took place more than a century ago, with cortex-to-cortex transplantations in experimental animals (Dunnett and Björklund 1994). Almost nine decades later, the hallmark finding that allografts of embryonic DA-ergic tissue survive, grow, develop graft-host connectivity, and reverse experimentally-induced parkinsonism in the rat (Björklund and Stenevi 1979; Perlow et al. 1979) implied that if the brain can be repaired and neurologic deficits can be reversed in experimental animals, this should also be possible to achieve in the human disease.

The first attempts to restore the striatal DA deficiency in PD by means of cell replacement were performed in 1982 and 1983 using autologous adrenal chromaffine tissue (Backlund et al. 1985). The fact that the patient was his own source of tissue meant that the method was free from ethical concerns associated with use of human embryonic tissue. Furthermore, risks for immune reactions and graft rejection, and the need for long-term immunosuppressive treatment, were avoided. Autologous transplantations were applied by centers all over the world during the 1980s (Rehncrona 1997), for PD as well as in a few cases of other parkinsonian disorders, such as progressive supranuclear palsy (Ward-Smith and Berry 1990). Graft survival was generally poor (Kordower et al. 1998b) and clinical trials typically revealed only transient and limited clinical benefits (Rehncrona 1997), which, in combination with a high incidence of mortality and morbidity (Goetz et al. 1991), led to abandonment of this approach.

1.4.2. Clinical trials

Animal experiments implanting cell suspensions of DA-rich tissue from allogeneic embryonic ventral mesencephalon (VM) into the denervated stratum showed increasingly promising results and by the end of the 1980s, the first clinical trials were performed in patients with PD (Lindvall et al. 1989). The primary goal in these first clinical trials was to explore whether grafted embryonic DA neurons can survive and have functional effects in the human brain. Initial studies showed only modest symptomatic relief after transplantation and there was no clear demonstration of graft survival (Lindvall 1994; Lindvall et al. 1989). The first evidence for survival of DA-ergic neurons in the graft, associated with substantial clinical improvement, were provided in 1990 (Lindvall et al. 1990). Subsequent replications by various independent
investigators also showed that intrastriatal DA-ergic grafts can survive and induce clinically valuable improvements in patients with PD (Defer et al. 1996; Freeman et al. 1995; Hauser et al. 1999; Lindvall et al. 1992, 1994; Peschanski et al. 1994; Remy et al. 1995; Sawle et al. 1992; Wenning et al. 1997). After a delay of at least 2-3 months, a majority of patients with surviving grafts gradually, over about 6 to 24 months, experienced a greater amount of time spent in the “on” phase, and reduced severity of rigidity and bradykinesia in the “off” phase. These trials have also shown that unilateral grafts can induce bilateral improvements (Defer et al. 1996; Lindvall et al. 1990, 1992, 1994; Wenning et al. 1997), and indicate that the functional recovery following transplantation, contra- as well as ipsilaterally to the graft, is incomplete both in terms of magnitude and pattern of improvement.

1.4.3. Assessment of graft survival and function

Survival of grafted DA-rich tissue in the host brain is a fundamental key factor for cell replacement therapy to induce substantial clinical recovery in PD (Hagell and Brundin 2001). Survival of grafted DA cells in transplanted patients is assessed in vivo by measuring striatal FD uptake on PET (Brooks 2000; Langston et al. 1992). Support for the validity of FD PET as a measure of striatal DA-ergic nerve function is strong (see above, section 1.2.2.1.). Additional support for the validity of this approach in relation to DA cell replacement therapy has been obtained from histopathological studies showing that the increased FD uptake on PET reflects substantial survival of DA-ergic grafts reinnervating the striatum in patients with clinically significant benefits (Kordower et al. 1995, 1996, 1998a).

The objective of neural transplantation in PD is to reverse clinical consequences of the disease by correcting a specific deficit in brain function, namely the impaired DA-ergic neurotransmission at denervated synaptic sites in the striatum. Whereas grafts of embryonic nigral neurons reinnervate the striatum, form synaptic connections, release DA, and improve motor deficits in experimental animals (Annett 1994; Brundin et al. 1994), in vivo graft function has not been clarified in patients.

1.4.4. Assessment of graft effects

The majority of clinical trials has been evaluated according to the Core Assessment Program for Intrastriatal Transplantations (CAPIT) (Langston et al. 1992). The purpose of CAPIT was to establish international consensus on the assessment of PD patients enrolled in clinical cell transplantation protocols, so that results from different clinical trials could be compared. The aim was to make CAPIT comprehensive as well as simple, accessible, and easy for anyone to use. Since then, CAPIT has been applied to
the assessment of patients not only in transplantation trials, but also in trials of other therapeutic interventions for PD, such as pallidotomy (Baron et al. 1996) and deep-brain stimulation (DBS) (Kumar et al. 1998). In addition to clinical assessments and brain imaging techniques (see above, section 1.4.3.), CAPIT includes recommendations regarding patient selection criteria as well as the graft tissue. These will, however, not be considered here.

1.4.4.1. Parkinsonism
The clinical assessment section of CAPIT comprises patient-derived “on/off” diaries to assess effects regarding daily motor fluctuations (see above, section 1.2.2.3.), and motor symptom assessments during single-dose L-dopa drug challenges (Langston et al. 1992). The single-dose L-dopa tests include assessments in the morning, before intake of the first anti-parkinsonian medication, and continual monitoring of motor function following the intake of an individually standardized single dose of L-dopa (preferably equal to the patient’s first regular dose of the day), until the patient has switched “off” again. See section 3.3.1. for more detailed descriptions.

1.4.4.2. Dyskinesias
The influence of grafts on dyskinesias in PD patients has been unclear and reported data are partly contradictory. One reason for this inconsistency probably relates to differences between transplantation procedures among investigators. Another may be the fact that few studies have included dyskinesias as a specific outcome. Insufficient attention to dyskinesia assessment in CAPIT (see above 1.2.2.3.; Langston et al. 1992) may, at least in part, have accounted for this. Clinical transplantation trials have generally thus assessed dyskinesias merely in terms of their daily duration, as derived from patients’ “on”/”off” diaries, or relied on clinical observations reported in case summaries. One exception is a study by Widner et al. (1992), where two patients with 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine-(MPTP)-induced parkinsonism were assessed using the AIMS while performing standardized motor tests during L-dopa challenges. Clear and progressive reductions in the amount of dyskinesias elicited by a single dose of L-dopa were evident up to two years following bilateral intrastriatal implantation of human embryonic VM.

1.4.4.3. Perceived health, activity limitations and participation restrictions
CAPIT does not provide any recommendations or suggestions regarding assessments beyond the level of impairment. Although section II of the UPDRS (“activities of daily living”) is included, this scale consists of a mixture of impairments and activity limitations and is, in its current version, of very limited value for this purpose (van Hilten et al. 1994). While it may be inferred from decreases in the time spent in “off”
that improvements have been evident also regarding, e.g., activity limitations and participation restrictions from the patients’ perspective, the validity of such assumptions is yet to be demonstrated. Furthermore, patients’ perceived PD-related problems are not limited to motor manifestations (Abudi et al. 1997; Brod et al. 1998; Frazier 2000), and non-motor features of the disease are common also in patients without motor fluctuations (Larsen et al. 2000). Given that the biological substrate(s) for the vast majority of such aspects of PD are largely unknown, and that cell replacement therapy is aimed to restore a specific neurobiological deficit within the brain, thorough evaluation of the effects of such interventions regarding various aspects of patients’ perceived health, disability, and illness-related distress is of great clinical and biological interest.
2. OBJECTIVES.

The overall aim of this explorative study was to apply, develop, and evaluate methods for assessment of dopamine cell replacement therapy in Parkinson’s disease at three different levels: body function (graft-derived restoration of striatal dopamine release), impairment (graft-induced effects on parkinsonism and dyskinesias), and patients’ perceived health. Specific objectives were:

- To apply an established protocol for clinical assessment of motor function in order to determine graft effects on parkinsonian symptomatology, in particular bradykinesia, following sequential bilateral transplantation in patients with Parkinson’s disease.

- To explore the possibility to assess the function of dopamine-rich grafts of human embryonic mesencephalic tissue, implanted into the striatum of patients with Parkinson’s disease.

- To develop a clinical rating scale for dyskinesias, at the level of impairment, that is easy to use and allows for detailed assessment of various types of dyskinesias, and to apply this scale to explore the effects of dopamine-rich intrastriatal grafts on dyskinesias in patients with Parkinson’s disease.

- To explore the influence of striatal dopamine cell replacement therapy on patients’ perceived health status and illness-related distress, and the value of health status questionnaires in the assessment of restorative interventions in Parkinson’s disease.

- To evaluate and compare the feasibility, psychometric properties, and dimensionality of generic and disease-specific health status questionnaires in Parkinson’s disease in order to assess their usefulness as outcome measures in restorative cell replacement interventions and to guide further development and evaluations.
Statistics are like a bikini: what is revealed is interesting; what is concealed is crucial

R. Taylor
3. METHODS

3.1. Ethical approvals

Patients’ consent was obtained according to the declaration of Helsinki and the procedures were approved by the local Research Ethical Committees in Lund, London, and Munich. Approval to administer radiolabel ligands was obtained from the Administration of Radioactive Substances Advisory Committee of the United Kingdom.

3.2. Grafting procedures

3.2.1. Patients

Patients were selected according to the CAPIT protocol (Langston et al. 1992). At study entry, all patients fulfilled the United Kingdom Parkinson’s Disease Brain Bank criteria for idiopathic PD (Gibb and Lees 1988). Basic preoperative clinical characteristics are summarized in Table 3.

<table>
<thead>
<tr>
<th>Patient no.</th>
<th>Age (years)</th>
<th>Disease duration (years)</th>
<th>Hoehn &amp; Yahr stage</th>
<th>Daily L-dopa dose (mg)</th>
<th>Implantation sites (no. of implant sites)</th>
<th>No. of donors</th>
<th>Follow-up (months)</th>
<th>Paper no.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>48</td>
<td>14</td>
<td>V</td>
<td>1200</td>
<td>L put + CN (2 + 1)</td>
<td>4</td>
<td>18</td>
<td>IV</td>
</tr>
<tr>
<td>2</td>
<td>55</td>
<td>14</td>
<td>IV</td>
<td>350</td>
<td>R put + CN (2 + 1)</td>
<td>4</td>
<td>17</td>
<td>IV</td>
</tr>
<tr>
<td>3</td>
<td>48</td>
<td>12</td>
<td>III</td>
<td>700</td>
<td>L put (3) / R put (5)</td>
<td>4 / 5</td>
<td>24 / 80</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15 / 71</td>
<td>IV</td>
</tr>
<tr>
<td>4</td>
<td>58</td>
<td>9</td>
<td>III</td>
<td>450</td>
<td>R put (3)</td>
<td>4</td>
<td>120</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>132</td>
<td>IV</td>
</tr>
<tr>
<td>7</td>
<td>49</td>
<td>10</td>
<td>II</td>
<td>200</td>
<td>L put (5) / R put (5)</td>
<td>5 / 5</td>
<td>24 / 41</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>73 / 56</td>
<td>IV</td>
</tr>
<tr>
<td>8</td>
<td>43</td>
<td>5</td>
<td>IV</td>
<td>500</td>
<td>L put + CN (4 + 2) / R put + CN (5 + 2)</td>
<td>5 / 8</td>
<td>18 / 41</td>
<td>I</td>
</tr>
<tr>
<td>9</td>
<td>52</td>
<td>10</td>
<td>IV</td>
<td>950</td>
<td>L put (5) / R put (5)</td>
<td>4 / 7</td>
<td>18 / 29</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16 / 27</td>
<td>IV</td>
</tr>
<tr>
<td>10</td>
<td>43</td>
<td>9</td>
<td>IV</td>
<td>850</td>
<td>R put + CN (5 + 2) / L put (5)</td>
<td>7 / 4</td>
<td>24 / 34</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50 / 60</td>
<td>IV</td>
</tr>
<tr>
<td>12</td>
<td>53</td>
<td>15</td>
<td>V</td>
<td>1,500</td>
<td>R put + CN (5 + 2) / L put + CN (5 + 2)</td>
<td>3 / 4</td>
<td>56</td>
<td>IV</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24</td>
<td>V</td>
</tr>
<tr>
<td>13</td>
<td>41</td>
<td>13</td>
<td>III-IV</td>
<td>225</td>
<td>L put + CN (5 + 2) / R put + CN (5 + 2)</td>
<td>4 / 5</td>
<td>20 / 26</td>
<td>V</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50 / 44</td>
<td>VI</td>
</tr>
<tr>
<td>14</td>
<td>49</td>
<td>11</td>
<td>III</td>
<td>425</td>
<td>R put + CN (5 + 2) / L put + CN (5 + 2)</td>
<td>4 / 4</td>
<td>23 / 24</td>
<td>V</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>56 / 55</td>
<td>IV</td>
</tr>
<tr>
<td>15</td>
<td>54</td>
<td>12</td>
<td>III</td>
<td>900</td>
<td>L put + CN (5 + 2) / R put + CN (5 + 2)</td>
<td>4 / 4</td>
<td>22 / 22</td>
<td>V</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>51 / 51</td>
<td>IV</td>
</tr>
<tr>
<td>16</td>
<td>68</td>
<td>12</td>
<td>III</td>
<td>1,075</td>
<td>R put + CN (5 + 2) / L put + CN (5 + 2)</td>
<td>4 / 3</td>
<td>18 / 18</td>
<td>V</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>48 / 48</td>
<td>IV</td>
</tr>
<tr>
<td>17</td>
<td>60</td>
<td>16</td>
<td>III</td>
<td>900</td>
<td>L put (7) / R put (7)</td>
<td>4 / 5</td>
<td>23 / 23</td>
<td>IV</td>
</tr>
<tr>
<td>18</td>
<td>51</td>
<td>10</td>
<td>V</td>
<td>1,250</td>
<td>R put (7)</td>
<td>4</td>
<td>19</td>
<td>IV</td>
</tr>
</tbody>
</table>

a Patients 5, 6, and 11 (MPTP-induced parkinsonism) are not included in the present study.
b For the first / second transplantation.
L, left; R, right; put, putamen; CN, caudate nucleus.
3.2.2. Tissue retrieval, procurement, and implantation

3.2.2.1. Paper I

Patients had received unilateral grafts either in the putamen alone or in the putamen plus caudate nucleus 10 to 56 months prior to the transplantations performed for this study (Table 3; Wenning et al. 1997). For detailed descriptions of tissue retrieval, procurement, and implantation, see Lindvall et al. (1989), Rehncrona (1997), and Wenning et al. (1997).

Human embryos were collected from routine elective abortions using ultrasound-guided suction technique. Immediately following abortion, the embryos were transferred into a Petri dish and transported to the laboratory for dissociation. Following rinsing, the dissected embryonic brain stem was subjected to four additional rinses under sterile conditions. Under microscopy, the VM was dissected out from the brain stem and cut into 6 to 10 small pieces that were stored in a hibernation medium (Sauer and Brundin 1991) up to 5 h during the retrieval and dissection of remaining donor tissue. The pieces were then incubated at 37°C for 20 min in Hank’s balanced salt solution (HBSS; Life Technologies) with 0.1% trypsin (Worthington), and 0.05% deoxyribonuclease (DNase; Sigma). After incubation, the pieces were rinsed five times using HBSS with 0.05% DNase. Immediately before implantation, the pieces were mechanically dissociated in HBSS with 0.05% DNase using fire-polished Pasteur pipettes.

Embryonic mesencephalic tissue from 4 to 8 aborted human embryos (aged 6 to 8 weeks post-conception; crown to rump length of 13 to 27 mm) was stereotaxically implanted to the non-operated side using a 1.0-mm outer diameter implantation cannula guided by preoperative computerized tomography (CT) and a stereotaxic frame attached to the patient’s skull. In all patients (numbers 3, 7-10 in Table 3), grafts were placed in the putamen along five trajectories in the ventrodorsal direction. Patient 8 was also grafted similarly along two trajectories in the head of the caudate nucleus on the same side. Twenty microliters of cell suspension were deposited in each putaminal trajectory, each of which was 12 to 14 mm long. The trajectories were planned to optimize reinnervation. Maximal distance between adjacent transplantation sites was 5 to 8 mm in the antero-posterior direction, the distance between the most anterior and posterior sites ranging from 22 to 29 mm. Caudate trajectories were 14 mm long and placed 4 mm apart.

All patients had received immunosuppressive therapy from two days before the first transplantation (Wenning et al. 1997) using a standard regimen of cyclosporine, azathioprine, and prednisolone (Lindvall et al. 1989). Apart from patients 3 and 8 (azathioprine discontinued at 20 and 6 months, respectively), this regimen was kept constant throughout the follow-up period.
3.2.2.2. Paper II
Procedures followed those outlined above (section 3.2.2.1.). In this patient (number 4; Table 3), dissociated ventral mesencephalic tissue from four human embryos (aged 6 to 7 weeks post-conception) was implanted unilaterally along three trajectories in the anterior, middle and posterior part of the right putamen. Immunosuppressive therapy was given for 64 months after surgery.

3.2.2.3. Paper IV
See above (section 3.2.2.1.) and below (section 3.2.2.4.) for details on patients 3, 4, 7-10, and patients 12-16, respectively. Patients 1 and 2 (Table 3; Lindval et al. 1989) were grafted with tissue from 4 donors each (aged 7 to 9 weeks post-conception) using a similar technique as in patients 3, 4, and 7-10, but with three differences: the implantation cannula was wider (outer diameter, 2.5 mm); saline rather than a balanced pH-stable salt solution was used during cell dissociation and storage; and the technique for loading the implantation cannula was less sophisticated, so that not all tissue could be used. In patients 17 and 18 (Table 3), the tissue was procured as described for patients 12-16 (section 3.2.2.4.), but was collected on three occasions during 7 days, and stored in a hibernation medium containing tirilazad mesylate and glial cell line-derived neurotrophic factor (GDNF) for 1 to 8 days prior to implantation (Petersén et al. 2000). Grafts were placed in the putamen along seven trajectories. Patient 17 was grafted bilaterally in the putamen using tissue from 4 and 5 donors (aged 5 to 9 weeks post-conception) on the left and right side, respectively. The second transplantation was performed four weeks following the first one. Patient 18 was grafted unilaterally in the right putamen, using tissue from 4 donors (aged 7 to 9 weeks post-conception). All patients received immunosuppressive therapy as described above (section 3.2.2.1.).

3.2.2.4. Paper V
Dissociated VM tissue from 7 to 9 aborted human embryos (aged 5 to 7 weeks post-conception; crown to rump length of 13 to 27 mm) was implanted bilaterally into the putamen and caudate nucleus in each of patient numbers 12-16 using CT- or magnetic resonance imaging (MRI) guided stereotaxic neurosurgery (Table 3). Tissue procurement followed that outlined above (section 3.2.2.1.), but with addition of the lazaroid tirilazad mesylate (Freedox®; Pharmacia Upjohn), a lipid peroxidation inhibitor, to the tissue rinsing, storage, incubation, and dissociation procedures in order to enhance graft survival (for detailed procedure description, see Brundin et al. 2000).

Tirilazad mesylate was also given intravenously to the patients four times a day (1.5 mg/kg) for 3 days, starting peroperatively at the time of the first implantation. Patients 12 and 16 were grafted bilaterally in one session, whereas patients 14 and 15 were operated with an interval of 4 and 2 weeks, respectively, between the two sides. Patient
13 received his second graft six months after the first one. Apart from patient 14 (azathioprine discontinued during the first month due to liver reaction), immunosuppressive therapy was given as described above (section 3.2.2.1.) for 12-24 months after surgery.

3.3. Assessments in grafted patients

3.3.1. Motor function

Patients were assessed according to the CAPIT protocol (Table 4; Langston et al. 1992) and according to a very similar protocol before CAPIT was introduced (Lindvall et al. 1989). Clinical evaluations were performed and videotaped in the practically defined “off” phase (i.e., in the morning ≥ 12 h after the last dose of anti-parkinsonian medication and ≥ 1 h after arising) and at regular intervals following the intake of an individually standardized single dose of L-dopa, which was the same at each assessment. Patients were assessed on no less than three occasions during six to twelve months preoperatively and four to twelve times annually for up to 132 months following transplantation (Table 3; Brundin et al. 2000; Hoffer et al. 1992; Lindvall et al. 1989, 1990, 1992, 1994; Wenning et al. 1997; papers I and II).

Motor function was assessed by means of timed motor tests and the UPDRS III (see above, section 1.2.2.2.) in the practically defined “off” and best “on” conditions. Timed motor tests and assessment of rigidity (according to the UPDRS or Lindvall et al. 1989) were also performed every 15-20 minutes throughout the L-dopa effect. Timed

<table>
<thead>
<tr>
<th>Table 4 Summary of clinical assessments according to the Core Assessment Program for Intracerebral Transplantations (CAPIT).</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A. Brain Imaging:</strong> MRI</td>
</tr>
<tr>
<td>FD PET</td>
</tr>
<tr>
<td><strong>B. Clinical Evaluation:</strong> &quot;Off&quot;/&quot;on&quot; diary</td>
</tr>
<tr>
<td>Single-dose L-dopa test</td>
</tr>
</tbody>
</table>

* The CAPIT protocol also includes recommendations regarding inclusion criteria and the graft tissue. These are not considered here.

* Practically defined "off" is defined as the condition in the morning ≥12 h after the last intake of antiparkinsonian medications and ≥1 h after arising.

MRI, magnetic resonance imaging; FD, 6-L-[18]F-fluorodopa; PET, positron emission tomography; preop, preoperatively; postop, postoperatively; L-dopa, levodopa given with a peripheral dopa-decarboxylase inhibitor; UPDRS, Unified Parkinson’s Disease Rating Scale.
tests measures the time (in seconds) required to: (a) perform 20 successive pronations/supinations of the hands with the patient sitting in a chair, tapping first the palm and then the back of his or her hand against the ipsilateral thigh; (b) tap the thumb with each finger, starting with the index finger and back again, 10 times (finger dexterity test); (c) tap the index finger back and forth 20 times between two points placed 30 cm apart on a table in front of the patient; and (d) walk 7 meters, turn around and walk back again. For detailed descriptions of test performance, see Hagell (2000). In addition, a subset of patients underwent neurophysiological measurements of upper limb velocity (simple arm and hand movements) in the practically defined “off” phase before and up to 24 months following final transplantation. These measurements were performed at the Medical Research Council Human Movement and Balance Unit, National Hospital for Neurology and Neurosurgery, London, UK (see Lindvall et al. 1989 for details).

3.3.2. Scanning procedures

All patients underwent FD PET scanning within 12 months before the first transplantation and, typically, during the second halves of the first and second postoperative years. Longer term follow-up scans were performed in patients 4, 7, and 12-16. All scans were performed after an overnight withdrawal of all antiparkinsonian medications, at the Medical Research Council Cyclotron Unit, Hammersmith Hospital, London, UK. Details on the FD scanning procedure are provided elsewhere (Sawle et al. 1992). FD uptake data from healthy control subjects were obtained for comparative purposes.

In patient number 4, we also performed PET scans using RAC (paper II). Postsynaptic striatal DA D2 receptor RAC occupancy was measured in the patient and five healthy control subjects after intravenous RAC injection. Each subject was scanned twice and was given an intravenous dose of saline in one scan and metamphetamine (0.3 mg/kg) in the other scan. Subjects were blinded to mode of injection. The PD patient stopped medication for at least 12 h before scanning.

3.3.3. Dyskinesias

Dyskinesias were assessed retrospectively in all patients (Table 3; paper IV) from video recordings during the practically defined “off” phase and at the peak of the L-dopa-induced “on” response. Ratings of hyperkinesias and dystonia were performed retrospectively from available patient videos using the Clinical Dyskinesia Rating Scale (CDRS; for details see below, section 3.4.) with a non-defined scoring code (maximum hyperkinesias and dystonia scores = 28; paper III). Observed and dominating type(s) of dyskinesias were also recorded. Videotapes were blinded for their dates of recording.
and rated in random order. The latest available preoperative recording, videos from about 12 and 24 months after grafting, and (for patients followed beyond that) the last available postoperative recording, were rated. Preoperative video recordings during practically defined “off” and peak “on” could not be located in three and two patients, respectively.

At the outset of the study, intrarater reliability was evaluated by means of ICC from repeated ratings of a separate set of 23 video sequences, used in the original evaluation of the scale (see below, section 3.4.; paper III), performed with an interval of three weeks. For ratings of hyperkinesias, ICC was 0.98 and for dystonia ratings 0.88. For purposes of this study, the overall amount of dyskinesias in each patient was also expressed using a global CDRS score, derived as the sum of the highest hyperkinesia or dystonia ratings from each body part (maximum score = 28; see below, section 3.4.2.). The ICC intrarater reliability coefficient for the global CDRS score was 0.98.

### 3.3.4. Perceived health and distress

Patients’ perceived health status was assessed in patients 12-16 using the NHP (Hunt et al. 1980; Hunt and McKenna 1989). The NHP is a generic health status questionnaire consisting of two parts. Part I includes 38 dichotomous (“yes”/”no”) items covering six scales: emotional reactions (ER; 9 items), sleep (SL; 5 items), energy (EN; 3 items), pain (PA; 8 items), physical mobility (PM; 8 items), and social isolation (SI; 5 items). Part II, which is optional and independent of part I, has seven questions covering seven domains of daily life. Part II has not proven very valuable and it is recommended that it no longer should be used (Bowling 1997b). Therefore, only NHP part I (hereafter referred to as the NHP) was used. Originally, each item of the NHP was assigned a weighted score (Hunt 1989). It has been shown, however, that such practice has no advantages (Jenkinson 1991; Prieto et al. 1996) and weights have not been developed for the more recent language versions of the NHP and were not applied here. Each NHP scale is scored from 0 to 100, where 100 indicates the worst possible health status. All items within each scale must be answered for calculation of scale scores (Hunt and McKenna 1989; Wiklund 1989).

Embedded in the NHP is the NHP index of Distress (NHPD), a 24-item measure of distress related to ill health, specifically omitting items relating to physical disability (McKenna et al. 1993). The NHPD has so far been sparsely used and evaluated (Martínez-Martín et al. 1999; McKenna et al. 1993; Whalley 1996). The NHPD can be administered alone or be derived from the full NHP and yields a score ranging between 0 and 24, where a score of 24 indicates the highest possible level of distress. Score calculation requires that no more than four item responses are missing (S.P. McKenna, personal communication).
The questionnaire was administered to patient numbers 12 through 16, who completed it independently at home, 1-3 months preoperatively and 5-7, 12 and 18-24 months postoperatively.

3.4. Development and evaluation of the Clinical Dyskinesia Rating Scale

3.4.1. Development of the scale

The Clinical Dyskinesia Rating Scale (CDRS) was developed with the purpose to fulfill the following criteria: (a) easy to use and apply to any situation, e.g., for multiple assessments during a drug-cycle while performing standardized motor tests for parkinsonism; (b) separate ratings of different body parts, including lateralization; (c) separate ratings of hyperkinesias and dystonia; and, (d) ratings should be at the level of impairment.

After explorative use in unselected parkinsonian patients and ratings of dyskinesias in squirrel monkeys with experimental MPTP-induced parkinsonism (Strandberg and Larsson 1995), the scale was modified to the versions used in this study.

3.4.2. Evaluation

The CDRS was evaluated regarding intra- and interrater reliability from a videotape containing a set of standardized sequences of PD patients displaying various types and severities of dyskinesias. The video sequences were rated according to the CDRS by a total of 13 clinicians experienced in PD. Details of the evaluation procedure are provided below.

3.4.2.1. The rating scale

During each scoring sequence, the highest severities of hyperkinesias and dystonia observed in the face (including tongue), neck, trunk, and left and right upper and lower extremities were scored. In a first evaluation, scores were assigned according to a non-defined 5-grade severity code (Table 5). In a second evaluation, ratings were performed according to a defined 5-grade severity code with definitions partly adopted from a suggested modification of the Rush dyskinesia scale (Goetz et al. 1994) (Table 5). For the second evaluation, the dystonia section was also clarified to cover only dystonic postures (brief or sustained). Both versions of the scale give a maximum total score of 28 for hyperkinesias and dystonia, respectively. Scores are assigned according to the highest severity observed during the rated sequence. Ratings are to be based on observations of the patient during activation and at rest. The scores are assigned to the body part where the dyskinesias are observed, e.g., hemiballism in an arm is assigned
to the arm, and face dyskinesias (e.g., tongue movements, blepharospasm or grimaces) are assigned to the face. The exception is abnormal involuntary head movements, which are assigned to the neck. The examiner may also add more detailed information on, e.g., the types of dyskinesias (preferably with identification of the dominating type), their relation to activation, or distinguishable lateralizations of axial dyskinesias (i.e., dyskinesias in the face, neck, and trunk) in a comments-section of the scale. A suggested form for rating according to the CDRS is given in paper III. In the original publication (paper III), the possibility to obtain an overall dyskinesia score using the CDRS was mentioned but never further evaluated. For the purpose of this thesis, a global CDRS score, derived as the sum of the highest hyperkinesia or dystonia ratings from each body part (maximum score = 28), was evaluated using data collected for paper III (see below, section 3.6.2.1.).

### Table 5 Rating code definitions for the Clinical Dyskinesia Rating Scale (paper III)

<table>
<thead>
<tr>
<th>Score</th>
<th>Non-defined</th>
<th>Defined</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>None</td>
<td>None observed</td>
</tr>
<tr>
<td>1</td>
<td>Mild</td>
<td>No interference with voluntary motor acts involved in the rated task</td>
</tr>
<tr>
<td>2</td>
<td>Moderate</td>
<td>There is interference with voluntary motor acts involved in the rated task, but it can be completed</td>
</tr>
<tr>
<td>3</td>
<td>Severe</td>
<td>There is intense interference with the voluntary motor acts involved in the rated task, and completion is greatly limited</td>
</tr>
<tr>
<td>4</td>
<td>Extreme</td>
<td>No completion of the voluntary motor acts involved in the rated task is possible</td>
</tr>
</tbody>
</table>

3.4.2.2. The video tape

Twenty-three sequences of a routine battery of standardized clinical motor tests performed during single-dose L-dopa-tests (Langston et al. 1992; Lindvall et al. 1989) in two patients with idiopathic PD (Gibb and Lees 1988) were edited to a 45 minutes long videotape. The patients displayed various types and severities of dyskinesias throughout their drug cycles. Each video sequence consisted of (a) the patient at rest, silent as well as speaking; (b) rigidity testing of the patients’ wrists, elbows and knees, performed by an examiner with the patient sitting in a chair; (c) bilateral performance of 20 pronations/supinations of the hands against the ipsilateral thigh with the patient sitting in a chair; and (d) the patient standing up, walking 7 meters, turning around and walking back again. The initial sequence of each patient also featured bilateral performance of 20 fist clenches, 10 finger dexterity movement cycles, and 20 foot taps with the patient sitting in a chair.
3.4.2.3. Raters and rating procedure
The raters consisted of eight neurologists, two neurosurgeons and three PD specialized nurses (two from a neurological and one from a neurosurgical clinic) from six different centra in three countries, all with longstanding clinical experience from PD. Six of the raters (raters I-VI) had not performed any formal clinical dyskinesia ratings in PD before and none of them were familiar with the CDRS prior to the evaluation. Raters were presented with the rating scale and standardized instructions consisting of the information provided above (sections 3.4.2.1. and 3.4.2.2.).

For the first evaluation, raters were instructed to rate the highest severity of dystonia and hyperkinesias observed in each body part separately for each of the 23 video sequences according to the non-defined 0-4 severity code (Table 5). The videotape was independently rated by each of the raters. Raters I-VI repeated the procedure two to three weeks after the first rating session, using the same videotape. In the second evaluation, ratings were performed using the defined version of the scale (Table 5). The ratings for the second evaluation were performed once by raters III, IV, VI and XI-XIII, using the same videotape as in the first evaluation.

In the first evaluation, raters were also scoring their global impressions of the levels of dyskinesias and parkinsonism in each rated video sequence (data not included in paper III). Global impressions were rated as none (= 0), mild (= 1), moderate (= 2), severe (= 3), or extreme (= 4).

3.5. Generic and disease-specific health status measurement in Parkinson’s disease

3.5.1. Design

This study was conducted as two parallel pilot studies, in this section referred to as studies one and two, respectively.

Study one was a clinical study conducted to evaluate content and linguistic validity of the disease-specific health status questionnaire PDQ-39 (see below, section 3.5.3.4.), which has undergone translation into Swedish but not been evaluated regarding these aspects. Study two was a cross-sectional postal survey designed to evaluate the feasibility and psychometric properties of the PDQ-39 and the generic NHP in PD.
3.5.2. Linguistic and content validity

3.5.2.1. Study sample
Study one involved nine consecutive non-demented patients with idiopathic PD (Gibb and Lees 1988) with a median HY stage of III (range: II-V), and three clinicians specialized in PD (one neurologist and two nurses).

3.5.2.2. Evaluation
Respondents were instructed to comment on each section and item as they carefully read the questionnaire (all respondents) and answered each item (patients). Next, respondents were asked a series of open ended questions: (1) Did you have any difficulty understanding or following the instructions?; (2) Did you have any problems with the choice of answers on the questionnaire?; (3) Did you feel there were any important issues related to health, function, and well-being that were missing from the questionnaire?; and (4) What was your overall impression of the questionnaire? For questions 1-3, respondents were also asked to rate their impression on a 1 - 10 scale, where 1 = “worst possible” and 10 = “best possible”. All comments and responses were recorded and reviewed by the patients at the end of each interview. Any difficulties observed by the investigator while patients responded to the questionnaire were also recorded.

Ratings of the questionnaire instructions, response alternatives, and content were interpreted according to the following arbitrarily predefined key: 1-4 = inferior; 5-7 = acceptable; 8-9 = good; 10 = excellent. In addition, patients’ responses to open ended questions 3 and 4 were used to evaluate the overall content validity of the PDQ-39.

3.5.3. Feasibility and psychometric properties

3.5.3.1. Study sample
For study two, every fifth patient (n = 81) was selected from a total of 407 PD patients receiving care at the Department of Neurology, Lund University Hospital, Sweden, during one year (1999). Participants in study one were excluded. Before selection for this study, the complete patient list was organized chronologically according to patients’ age in order to avoid a possible age bias.

3.5.3.2. Data collection
Health status questionnaires were mailed to the patients together with demographic questions and questions designed to obtain indices of PD severity and overall QoL, a
covering letter and an addressed, stamped response envelope. Questionnaire response was interpreted as consent to participate. No reminders were sent out.

3.5.3.3. Generic health status
The Swedish language version of the generic NHP (Wiklund 1989; see above, section 3.3.4.) was used. Responses were scored as described previously (section 3.3.4.) but in this study the raw scores of the NHPD were transformed (observed score / highest possible score x 100) to a possible range between 0 and 100 (100 = highest level of distress) in order to conform with other NHP scales and with the disease-specific questionnaire (see below, section 3.5.3.4.).

3.5.3.4. Disease-specific health status
The PDQ-39 was developed in the UK as a measure of functioning and well-being for patients with PD (Jenkinson et al. 1998; Peto et al. 1995) and consists of 39 items covering eight scales: mobility (MOB; 10 items), activities of daily living (ADL; 6 items), emotional well-being (EMO; 6 items), stigma (STI; 4 items), social support (SOC; 3 items), cognitions (COG; 4 items), communication (COM; 3 items), and bodily discomfort (BOD; 3 items). Respondents are requested to indicate how often, during the past month (“frame question”), they have experienced the problems defined by each item according to either of five response categories (“never” - “occasionally” - “sometimes” - “often” - “always, or cannot do at all”). Scale scores range between 0 and 100, where 0 = no problem, and 100 = maximum level of problem, and their calculation require all items within the scale to be answered (Jenkinson et al. 1998).

Based on procedures established by the International Quality of Life Assessment (IQOLA) group (Bullinger et al. 1998), the developers of the British source version have adopted the PDQ-39 into, among others, the Swedish language version (PDQ-39se) used in this study. Briefly, the procedure included forward translation by bilingual translators native in the target languages, followed by back translation into English by British bilingual translators. Reconciliation meetings were held between the translators and the developers of the British source version of the PDQ-39, where conceptual and semantic equivalence, as well as any translation differences and ambiguities, were discussed before the new language versions were agreed upon (D. Wild, personal communication).

3.5.3.5. Additional variables
Patients’ perceived severity of PD was assessed by a single question: “Overall, how do you perceive the severity of your PD?” Respondents were given three alternatives: “mild”, “moderate”, and “severe”, scored 1 through 3, respectively (Hagell and Sandlund 2000).
Patients’ perceived overall QoL was assessed by a single question: “Everything taken together, how do you perceive your overall quality of life?” Respondents were to indicate their response by placing a mark on a 10-cm horizontal visual analogue scale (VAS) anchored by “worst imaginable quality of life” at the far left, and “best imaginable quality of life” at the far right. Scores were derived as the distance, in millimeters, from the left end of the VAS, thus yielding a possible score between 0 - 100, where 100 indicates the best imaginable QoL.

Demographic questions included duration of PD, number of daily PD medicine intake occasions, and living conditions.

3.6. Analyses

3.6.1. General data analytic approach

Variables violating more than one of the assumptions underlying the use of parametric statistics, i.e., interval or ratio level data, normal distribution, and homoscedasticity, are described as median and inter-quartile range (IQR) and analyzed by means of non-parametric statistics. Other data are described as mean ± SD and analyzed by means of parametric statistics (Norman and Streiner 2000; Siegel and Castellan 1988). Hence, the Wilcoxon signed rank test and paired t-test are used for comparisons between two related samples; the Mann-Whitney U test and unpaired t-test are used for comparisons between two independent samples; the Friedman test is used for comparison between > 2 related samples; Kruskal-Wallis one-way analysis of variance (ANOVA) by ranks is used for comparisons across > 2 independent samples, and correlations among variables are explored by Spearman’s rank order correlation (r) and Pearson’s product-moment correlation (r). P-values are 2-tailed. The alpha level of significance is 0.05. Statistical analyses were performed using SPSS 10.1 for Windows (SPSS Inc., Chicago, IL, 2000) and StatView 5.0 for Macintosh (SAS Institute Inc., Cary, NC, 1998).

3.6.2. Specific analyses

Methods of analyses employed in specific parts of the current study are described below.

3.6.2.1. Evaluation of the dyskinesia rating scale

In paper III, Kendall’s rank-order correlation coefficient (T) (Siegel and Castellan 1988) was used for determination of intrarater consistency (first evaluation; raters I-
VI), comparing the first with the second sets of hyperkinesia and dystonia ratings. For
the analysis of interrater concordance, the Kendall coefficient of concordance ($W$) was
determined for the hyperkinesia and dystonia sections of the scale (first evaluation,
raters I-X; second evaluation, raters III, IV, VI, and XI-XIII), and the Chi-Square value
was obtained in order to determine statistical significance (Siegel and Castellan 1988).

Further analyses of data from the first six raters were performed for the purpose of this
thesis. Other raters were excluded because they either had used the defined rating
scale version and/or had not performed repeated ratings. First, inter- and intrarater
reliability were re-assessed by means of the ICC. This is because the ICC generally is
considered the measure of choice evaluation of test-retest, intra-, and interrater reliability
(see above, section 1.3.2.). Additionally, there are observations in the literature
indicating that Kendall’s $W$, used in the original evaluation of inter-rater reliability
(paper III), may be biased by the degree of similarity in evaluated scores (Henderson
et al. 1991). Second, the use of an overall dyskinesia score (the global CDRS score),
as suggested but not evaluated in paper III, was evaluated regarding reliability (ICC)
and relation to the raters’ global impression of dyskinesias and parkinsonism from the
respective video sequences. It was hypothesized that a valid global CDRS score should
correlate strongly and positively with the dyskinesia global impression score and
negatively with the global impression of parkinsonism.

3.6.2.2. Responsiveness of outcome measures
For the purpose of this thesis, the responsiveness of the various NHP-derived scales,
as well as of the UPDRS motor score, pronation/supination test, the percentage of
daily time spent in the “off” phase, and patients’ daily L-dopa requirements, as assessed
preoperatively and during the second postoperative year in patients assessed by these
means at the same time points before grafting and during the second postoperative
year (Brundin et al. 2000; paper V), were explored by calculation of their respective
ESs (change score / baseline SD).

3.6.2.3. Feasibility of the NHP and PDQ-39se
In paper VI, feasibility was assessed for each instrument as a whole as well as for the
individual questionnaire scales. Instruments were considered feasible if 70% or more
of respondents answered all items. Scales were considered feasible if the proportion
of responses precluding scoring was less than 15%. Thresholds were arbitrarily defined
for this study.

3.6.2.4. Psychometric properties of the NHP and PDQ-39se
In paper VI, internal consistency was assessed according to Cronbach’s coefficient
α (Cronbach 1951). The percentage of responses yielding the lowest and highest possible scores (floor and ceiling effects) were examined for each scale. The threshold for acceptable floor- and ceiling effects was set at 20% (Holmes and Shea 1997; McHorney et al. 1994). Known-groups validity was assessed across patients’ perceived PD severity using Kruskal-Wallis ANOVA. Scores were expected to indicate more impaired health status by increasing perceived PD severity. Convergent and divergent validity of the PDQ-39se and NHP was evaluated by a multitrait-multimethod matrix, and were considered supported when NHP and PDQ-39se scales of the same trait correlated stronger with one another (convergent validity) than with scales tapping other traits (divergent validity) (Fayers and Machin 2000). Because it has been implied that both the PDQ-39 and NHP measure QoL (Martínez-Martín et al. 1999; Schrag et al. 2000), we also explored this possibility by assessing the Spearman correlations between scores on these scales and patients’ perceived overall QoL.

Rasch analyses were performed using the Rasch measurement software WINSTEPS, version 3.24 for Windows (MESA Press, Chicago, IL, USA; 2001). NHP derived scales were analyzed by the dichotomous Rasch measurement model (Wright and Stone 1979). PDQ-39se scales were analyzed using the Rasch partial credit model, which assesses each item’s response categories independently from the other items (Wright and Masters 1982), and thus allows exploring individual item response category step threshold measures (i.e., the logit measures that separate between response categories 0 to 1, 1 to 2, 2 to 3, and 3 to 4). Unidimensionality was determined by INFIT and OUTFIT standardized Z statistics, which indicate misfit and, thus, multidimensionality when below -2.0 or above +2.0 (Smith 2000; Smith et al. 1998). Item and response category step difficulties and hierarchies were determined by their logit measures. Separation indices were used to determine whether scale items separated respondents into levels of ability, expressed as the number of person strata, which represent the number of statistically distinct levels (separated by at least three errors of measurement) of person ability defined by each scale (Smith 2001; Wright and Masters 1982). Although two strata is considered a minimum (Smith 2001), clinically useful scales should preferably encompass at least three strata, e.g., mild, moderate, and severe (Tesio and Cantagallo 1998). The extent to which scales targeted the measured respondents (i.e., excluding those with floor- and ceiling scores) was assessed by examination of the respondent mean ability measure in relation to the mean item difficulty measure for each scale, predefined as 0 logits. Finally, we explored the presence of potential DIF between men and women, and younger and older respondents. Hence, NHP and NHPD item difficulty measures and PDQ-39se item response category step measures were calibrated separately for men (n = 44), women (n = 27), and two age groups (younger, n = 39; older, n = 32) identified by the median age of the whole sample. Calibrations were then compared between gender and age groups using a 95% confidence interval to identify potential DIF.
What we see depends mainly what we look for

John Lubbock
4. RESULTS

4.1. Motor effects of sequential bilateral DA-rich intrastriatal grafts (Paper I)

Five unilaterally grafted parkinsonian patients (patients 3 and 7-10; Table 3) were followed regarding graft survival and effects on motor functions for 18-24 months after a second, contralateral transplantation performed 10-56 months after the first one. All patients showed surviving grafts. During the second year after transplantation, putaminal FD uptake corresponding to the second graft was elevated by a mean of 85% from baseline, whereas the previously grafted, contralateral, putamen did not show any significant changes (an overall 7% decreased FD uptake). At the time of the postoperative PET scan performed 12-18 months after transplantation, the bilateral putamen FD uptake in this group of patients was 52% of the normal mean, compared with 36% before the first and 42% before the second transplantation.

The percent time spent in “off”, which decreased by about 50% in patients 3, 7 and 10 after the first transplantation, exhibited a further marked reduction in patients 7 and 10, in whom ”off” periods disappeared or became very brief. Patient 3 did not experience any persistent further decrease in the amount of time spent in the “off” phase, which, similarly to following the first implantation, remained around 30% of the day (compared to about 50% preoperatively). In agreement, no consistent alterations of the duration of the response to a single dose of L-dopa were observed in patient 3, which was in contrast to a lengthening by about 1 h after the first implant. Patients 7 and 10, on the other hand, exhibited marked prolongation of the L-dopa response after both transplantations.

Fig. 1 illustrates the overall development of parkinsonian symptomatology as assessed by the UPDRS motor score during practically defined “off”. Three patients had already experienced clinically useful symptomatic relief after the first graft (Wenning et al. 1997). Two of them (patients 7 and 10) showed further marked improvements whereas one patient (number 3), who received his second graft 56 months following the first one, also improved but not to the extent that he regained his motor status from the second year following the first graft. Two patients (numbers 8 and 9) who had only minor functional effects after the first implantation continued to deteriorate following the second transplantation. The clinical course of patient 9 was complicated by development of additional pathology, which may have compromised the beneficial effects of the second graft. Already before the second transplantation, he started exhibiting cognitive impairments and developed dementia during the follow-up period. Patient 8 exhibited an atypical clinical development, raising a suspicion that he does not suffer from idiopathic PD, but multiple system atrophy (Wenning et al. 1997). The virtual lack of symptomatic relief in this patient, whose FD uptake in the grafted putamen
reached levels similar to those in patients 3, 7 and 10, may be explained by the atypical nature of his parkinsonism. In order to evaluate the effects of the first and the second grafts separately, the time taken to perform 20 pronations/supinations with the hands was analyzed separately for each of the upper limbs during various intervals before and after the first and the second grafts. Results are depicted in Fig. 2. These data illustrate several aspects of graft effects on clinical motor performance. First, graft effects are evident bilaterally, but are more pronounced contralaterally to the graft. This pattern is seen following both the first and the second graft. Second, the magnitude of graft efficacy is lesser following the second than the first transplantation, ipsi- as well as contralaterally. In patients 3, 7, and 10, neurophysiologically derived measures
of movement time followed a similar pattern as that observed using the timed pronation/supination test. The correlation between pronation/supination tests and movement time, as derived from multiple test occasions preoperatively and throughout the follow-up, was moderate to strong \((r = 0.726, 0.766, \text{ and } 0.489 \text{ in patients 3, 7, and 10, respectively})\).

4.2. Monitoring graft function in vivo (paper II)

The possibility to monitor synaptic DA release was explored in a grafted PD patient with pronounced postoperative improvements following unilateral transplantation to the putamen 10 years earlier (patient 4 in our series; Table 3). The patient improved clinically up to three years after transplantation (Fig. 3a), by which time motor fluctuations, “off” phases and rigidity had disappeared and bradykinesia was markedly improved. Graft effects were evident bilaterally but were more prominent contralaterally to the grafted putamen (Wenning et al. 1997). L-dopa was withdrawn after 32 months and, due to slight symptom progression axially and ipsilaterally to the graft, reintroduced 3.5 years later using 1/3 of the preoperative dose. In parallel with clinical improvement, FD uptake in the grafted putamen increased up to three years post-surgery, by which time it reached normal levels, with only minor additional changes up to 10 years after transplantation, whereas FD uptake in non-grafted striatal regions decreased gradually (Fig. 3b).
We used PET and the DA D2 receptor ligand RAC to measure D2 receptor occupancy by endogenous DA in the presence and absence of metamphetamine. Compared with healthy subjects, RAC binding was upregulated in non-grafted striatal areas but normal in the grafted putamen in the baseline condition (Fig. 3c). Following metamphetamine administration, RAC binding was reduced by 4.5% in the non-grafted, whereas it decreased by 26.6% in the grafted putamen, which was similar to the decrease in healthy subjects (Fig. 3c). No side-to-side difference in the decrease of RAC binding induced by metamphetamine was observed between the caudate nuclei.

4.3. Development and application of a clinical dyskinesia rating scale (papers III & IV)

4.3.1. Reliability (paper III)

Based on experience from available rating scales, we devised a tool for quantification of dyskinesias in PD, the Clinical Dyskinesia Rating Scale (CDRS). The CDRS was designed to be easy to use, provide separate ratings for different body parts, as well as of hyperkinetic and dystonic dyskinesias, and to quantify dyskinesias at the level of impairment rather than activity limitations. Tests regarding inter- and intrarater reliability, using a non-defined rating code, yielded an overall intrarater consistency coefficient (T) of 0.738 and 0.741 for hyperkinesias (range: 0.636 - 0.904) and dystonia (range: 0.315 - 0.917), respectively (raters I-VI). Interrater concordance (W) for hyperkinesias ranged between 0.859 and 0.914 when using a non-defined scoring code (raters I-VI and I-X), and was 0.883 for the defined scoring code (raters III, IV, VI, ...
and XI-XIII), respectively. Corresponding values for dystonia ratings were 0.317 – 0.474 and 0.440. Scores obtained by the defined and non-defined scoring codes did not differ for either hyperkinesia or dystonia ratings (p = 0.401 and 0.488, respectively; Kruskal-Wallis ANOVA).

Results from additional analyses performed for the purpose of this thesis are presented in Table 6. High levels of intrarater reliability were observed for both hyperkinesias and dystonia ratings. This was the case for hyperkinesia ratings also regarding interrater reliability, whereas dystonia ratings did not display acceptable interrater reliability. The global CDRS score reached acceptable levels of both intra- and interrater reliability. Results regarding ratings of hyperkinesias and dystonia are in general agreement with those obtained using Kendall’s T and W in the original analyses of these data (paper III). However, the coefficient obtained for dystonia ratings was lower when using ICC as compared to Kendall’s W.

4.3.2. Assessment of the influence of intrastriatal grafts of human embryonic mesencephalic tissue on dyskinesias (paper IV)

Fourteen grafted PD patients were assessed and analyzed retrospectively regarding the presence, severity, and types of dyskinesias before and up to 11 years after grafting. All patients showed peak “on” dyskinesias before grafting. Of 11 cases for whom preoperative video recordings during “off” were available, six displayed some degree of “off” dyskinesias prior to transplantation. Except for very mild repetitive hyperkinetic movements of the right leg in one patient and mild, activity-induced choreiform neck movements in another, preoperative “off” phase dyskinesias were dystonic.

---

Table 6: Explorative correlation matrix and re-evaluations of intra- and interrater reliability of the CDRS and raters’ global impressions of degrees of dyskinesias and parkinsonism

<table>
<thead>
<tr>
<th></th>
<th>CDRS scores b</th>
<th>Global Impression d</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hyperkinesia</td>
<td>Dystonia c</td>
</tr>
<tr>
<td>CDRS score:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hyperkinesia</td>
<td>0.865 / 0.851</td>
<td></td>
</tr>
<tr>
<td>Dystonia</td>
<td>0.246</td>
<td>0.862 / 0.258</td>
</tr>
<tr>
<td>Global</td>
<td>0.977</td>
<td>0.332</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.925 / 0.814</td>
</tr>
<tr>
<td>Global Impression:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dyskinesias</td>
<td>0.833</td>
<td>0.450</td>
</tr>
<tr>
<td>Parkinsonism</td>
<td>-0.756</td>
<td>-0.518</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.781</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.771</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.864 / 0.806</td>
</tr>
</tbody>
</table>

a Correlations (Spearman’s Rho; all p-values are < 0.004, 2-tailed) are in the lower left triangle of the table and reliability coefficients (ICC; intrarater / interrater reliabilities, respectively) are in the diagonal. Based on data from the first rating performed by raters I-VI (paper III).

b Sum of ratings of observed dyskinesia severity in the upper and lower extremities, trunk, face, and neck, using a non-defined scoring code (0 = none; 4 = extreme); maximum score = 28.

c Sum of the highest hyperkinesia or dystonia ratings from each body part (maximum score = 28).

d Rated on a 0-4 scale (0 = none; 4 = extreme).

CDRS, Clinical Dyskinesia Rating Scale.
Hyperkinesias and dystonias increased significantly during “off” after transplantation (Fig. 4), both reaching their maximum at 39.8 (range: 11-132) months. At this point, the median “off” phase hyperkinesia and dystonia CDRS scores were 4.0 and 2.75, respectively, as compared to 0 and 0.5 preoperatively. Severity of peak “on” dyskinesias or the time spent in “on” with dyskinesias did not alter significantly. Postoperative “off” phase hyperkinesias and dystonia typically appeared concurrently, as choreiform movements intermingled with brief dystonic postures. Repetitive, stereotypic or ballistic movements were also observed.

Some degree of postoperative “off” phase dyskinesias was observed in all transplanted cases. In eight patients, the maximum postoperative “off” phase global CDRS scores ranged between 0.5 and 4.5 (median [IQR]: 2.5 [1.1-3.8]). In these patients, dyskinesias were mild, and patients rarely reported any distress or disability related to “off” phase dyskinesias. In several cases neither the patients nor their assessors had paid any particular previous attention to them. Six patients exhibited episodes of more pronounced “off” phase dyskinesias, ranging between 9 and 18 (median [IQR]: 12 [10.1-15.4]), causing some degree of disability and/or annoyance in all, and constituting a clinical therapeutic problem in one case. The maximum “off” phase global CDRS

![Fig. 4](image)

**Fig. 4** Hyperkinesias and dystonia during practically defined “off” and peak “on” phases before transplantation, at the time of maximum postoperative dyskinesias (mean 39.8 months), and at the latest assessment (mean 44.6 months). Solid horizontal lines are median values, boxes are inter-quartile ranges, error bars are ranges. Open and filled circles are outliers and extremes (>1.5 and >3 box lengths from the 25th/75th percentiles, respectively). Hyperkinesias and dystonia during “off” increased significantly over time (p < 0.0001 and < 0.05, respectively), whereas peak “on” dyskinesias were unchanged (Friedman test). Data from paper IV.
scores differed significantly between the two groups of patients (p = 0.002; Mann-Whitney U test). Also the “on” phase dyskinesias were more pronounced in the latter group at the time of the maximum postoperative “off” phase dyskinesias (median [IQR]: 13.8 [6.8-20.2] vs 5.2 [3-6.6], p = 0.020; Mann-Whitney U test), whereas there was no significant difference preoperatively.

Differential development of “off” and “on” dyskinesias was observed post-grafting. For example, two patients with virtually no preoperative “off” dyskinesias developed mild dyskinesias in “off”, whereas their pronounced preoperative “on” dyskinesias were reduced by more than 50%. Postoperative “off” dyskinesias, as defined by the global CDRS score, correlated negatively with preoperative putaminal FD uptake ($r_s = -0.549$, $p = 0.064$). No correlation was found between “off” phase motor improvement, as assessed using the UPDRS motor score, and the maximum postoperative ”off” CDRS scores ($r_s = 0.003$, $p = 0.991$). Furthermore, “off” dyskinesias typically developed at a later time point following grafting as compared to the relief of parkinsonian symptomatology.

No evidence was obtained for the notion that ”off” dyskinesias are caused by overgrowth of grafted DA neurons. Neither the FD uptake in the scan performed closest in time to maximum postoperative “off” phase dyskinesias, nor the increase of putaminal FD uptake as compared to the preoperative scan, correlated with “off” global CDRS scores ($r_s = -0.267$, $p = 0.401$). However, postoperative “off” CDRS scores correlated with the number of VM implanted in the putamen ($r_s = 0.562$; $p = 0.037$), and tended to correlate with the number of putamen trajectories ($r_s = 0.492$, $p = 0.074$). Two patients grafted with tissue that had been stored for 1 to 8 days in the presence of GDNF developed more pronounced “off” dyskinesias than patients implanted with fresh tissue.

### 4.4. Perceived health following bilateral DA-rich intrastriatal grafts (paper V)

In paper V, the generic health status questionnaire Nottingham Health Profile (NHP), including the NHP index of distress (NHPD), was used to explore the influence of intrastriatal DA-rich grafts on perceived health and illness-related distress, as well as the usefulness of such questionnaires in cell replacement therapy trials for PD. Five patients implanted with human embryonic VM tissue bilaterally in the putamen and caudate nucleus were assessed (patients 12-16; Brundin et al. 2000; Table 3).

All administered NHP-forms were completed and returned according to the protocol without any missing data. All patients except for number 13 demonstrated deficits within virtually all NHP scales preoperatively. Two of the patients (numbers 12 and 14) showed clear improvements within all preoperatively affected dimensions at 18-24 months postoperatively. With few exceptions, these changes were stable or increased throughout the follow-up period. Already preoperatively, patient 13 did not indicate
any problems according to the NHP except for within the PM scale, which improved postoperatively. Patient 15 improved within three NHP scales (PM, SI and ER) and remained stable within the other three. Patient 16 improved within two dimensions (EN and PM), deteriorated within one (SI), and was stable within the other two. All patients but number 13, who scored zero already preoperatively, improved on the NHPD. The mean NHPD score decreased from 7.6 preoperatively to 2.4 at 18-24 months after surgery (p = 0.068; Wilcoxon’s signed rank test). Statistically significant reductions for the whole group were found for the PM dimension (p<0.05; Wilcoxon’s signed rank test). No statistically significant changes compared to preoperatively were observed within the other dimensions of the NHP.

### Table 7 Effect sizes for various outcome measures as assessed up to 24 months following bilateral striatal implantation of embryonic mesencephalic tissue in Parkinson’s disease

<table>
<thead>
<tr>
<th></th>
<th>Change score a</th>
<th>SD preop</th>
<th>Effect size b</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHP scale:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ER</td>
<td>-26.6</td>
<td>24.1</td>
<td>1.10</td>
</tr>
<tr>
<td>SL</td>
<td>-20.0</td>
<td>32.9</td>
<td>0.61</td>
</tr>
<tr>
<td>EN</td>
<td>-40.0</td>
<td>38.0</td>
<td>1.05</td>
</tr>
<tr>
<td>PA</td>
<td>-12.5</td>
<td>16.8</td>
<td>0.74</td>
</tr>
<tr>
<td>PM</td>
<td>-25.0</td>
<td>19.0</td>
<td>1.32</td>
</tr>
<tr>
<td>SI</td>
<td>-4.0</td>
<td>20.0</td>
<td>0.20</td>
</tr>
<tr>
<td>NHPD</td>
<td>-5.2</td>
<td>5.2</td>
<td>1.00</td>
</tr>
<tr>
<td>UPDRS motor score c</td>
<td>-16.6</td>
<td>16.2</td>
<td>1.02</td>
</tr>
<tr>
<td>Daily % time &quot;off&quot; d</td>
<td>-12.8</td>
<td>19.1</td>
<td>0.67</td>
</tr>
<tr>
<td>Pronations/supinations c,e</td>
<td>Right</td>
<td>-4.3</td>
<td>8.4</td>
</tr>
<tr>
<td>Left</td>
<td>-30.8</td>
<td>61.8</td>
<td>0.50</td>
</tr>
<tr>
<td>Daily L-dopa requirement</td>
<td>-370.0</td>
<td>510.8</td>
<td>0.72</td>
</tr>
</tbody>
</table>

a Negative change scores indicate improvement.
b Change score divided by the preoperative SD. Effect sizes of 0.20 to 0.49 are small; 0.50 to 0.79, moderate; and ≥0.8, large (Kazis et al. 1989; McDowell and Newell 1996).
c As assessed in the practically defined "off" phase.
d As assessed from patients’ self-scored "on"/"off" diaries.
e Time to perform 20 pronation/supination movements with the hand.

SD, standard deviation; NHP, Nottingham Health Profile, ER, emotional reactions; SL, sleep; EN, energy; PA, pain; PM, physical mobility; SI, social isolation; NHPD, the NHP index of distress; UPDRS, Unified Parkinson’s Disease Rating Scale; L-dopa, levodopa given with a peripheral dopa-decarboxylase inhibitor.

Data from paper V and Brundin et al. 2000.
For purposes of this thesis, the responsiveness of the various NHP-derived scales, as well as of measures of parkinsonian motor symptoms, as assessed preoperatively and during the second postoperative year (Brundin et al. 2000), were explored by calculation of their respective ESs (Table 7). According to general interpretation guidelines (see above, section 1.3.4.), results indicate that most NHP scales display responsiveness at the moderate (SL and PA) to large (ER, EN, PM, and the NHPD) level, whereas only one (SI) displays a small effect size. These effect sizes are comparable to those observed for traditional clinical outcome measures, among which the UPDRS motor score displays a large effect size, whereas other effect sizes are at the moderate level (Table 7).

4.5. Evaluation and comparison of generic and disease-specific health status questionnaires in Parkinson’s disease (paper VI)

4.5.1. Linguistic and content validity of the Swedish PDQ-39

Because the Swedish language version of the disease-specific PDQ-39 (PDQ-39se) has not been evaluated regarding its linguistic and content validity, these aspects were addressed in a sample of nine PD patients and three clinicians specialized in PD.

The PDQ-39se caused problems or raised doubts due to linguistic and, in part, stylistic causes among patients and clinicians in several instances. Problems with the frame question (identified by seven patients and one clinician) related mainly to it appearing anonymous with patients not taking it into consideration when reading the items. The distinction between response alternatives, in particular between “sometimes” and “occasionally”, was found unclear by five patients and one clinician. Ratings of the response alternatives yielded, for patients, a median (range) of 8 (5-10), and for clinicians 9 (4-9). Several problems were apparent related to items 28 and 29, which ask about support from spouse or partner, and support from family and close friends, respectively. The main problem was a double negative in the wording, which even caused some patients to give the opposite answer to that intended. Other comments regarding these items related more to the content and included the separation of spouse/partner and family/friends, which was felt unnecessary by one patient, and six respondents considered item 28 irrelevant to those living alone.

Patients identified several aspects related to health, functioning, and well-being that they considered important but that were missing from the questionnaire, e.g., issues related to medication, nutritional aspects, dyskinesias, sexual problems, and unpredictability of symptoms. Nevertheless, most respondents felt that, overall, the questionnaire was comprehensive and addressed important aspects of the disease. The
content was rated as acceptable to good by both patients (median [range]: 8 [6-10]) and clinicians (9 [5-10]).

### 4.5.2. Feasibility and psychometric properties of the NHP and PDQ-39

The feasibility of the NHP (including the NHPD) and PDQ-39se was evaluated based on the response rates to a cross-sectional postal survey, where the questionnaires were mailed to 81 PD patients, and the psychometric properties of the scales were evaluated using traditional as well as Rasch measurement approaches.

#### 4.5.2.1. Feasibility

Seventy-one questionnaire packages were returned, yielding an overall response rate of 88%. Missing item responses were found in 22.5% and 25.3% of returned PDQ-39se and NHP questionnaires, respectively. In most instances, PDQ-39se scales had somewhat better response rates than those of the NHP. The percentage returned PDQ-39se scales that could not be scored was highest for the MOB and SOC scales (7% and 12.7%, respectively); others ranged between 0% and 4.2%. NHPD performed better than the traditional NHP scales (4.2% vs 7% to 15.5%). The ER scale (15.5% responses not possible to score) of the NHP was the only one failing to meet the preset 15% criteria.

#### 4.5.2.2. Traditional psychometric analyses

The SOC and BOD scales of the PDQ-39se, and the SL, EN, and SI scales of NHP failed to demonstrate acceptable levels of internal consistency (coefficient α: 0.63 - 0.78). Floor effects were evident within all NHP and three PDQ-39se scales. Only the EN scale of the NHP exhibited substantial ceiling effects. Floor effects decreased by increasing perceived PD severity and were substantial for all scales among respondents considering their PD as mild. Among those who perceived their PD as severe, these persisted only for the SOC (PDQ-39se), SL and SI (NHP). Known-groups validity was supported for all scales but the SOC (PDQ-39se) and SL (NHP), which both failed to demonstrate significant score variability by perceived PD severity (p > 0.05; Kruskal-Wallis ANOVA). Correlations between PDQ-39se and NHP scales scores tapping the same underlying constructs supported convergent and divergent validity for mobility (MOB and PM), emotional (EMO and ER), and pain (BOD and PA) scales, which all showed the strongest correlations with one another. Social scales did not show this pattern.

Correlations between PDQ-39se and NHP scales and respondents’ perceived overall QoL were typically within the moderate range (ranges \(r\): PDQ-39se scales, -0.202 to -0.600; NHP scales, -0.497 to -0.650) and were similar to those between overall QoL
and indices of PD severity: patients’ perceived PD severity \( r_s = -0.564 \), number of daily intakes of antiparkinsonian medications \( r_s = -0.450 \), and duration of disease \( r_s = -0.587 \). The exception was NHPD, which showed a stronger correlation with perceived QoL than the other scales \( r_s = -0.747 \). The average correlation for the traditional NHP scales (excluding NHPD) was -0.585, and for PDQ-39se it was -0.478. Negative coefficients are due to opposite scoring directions.

4.5.2.3. Rasch measurement analyses

Scale items’ relative difficulties and hierarchies are depicted in Fig. 5, which also indicates misfitting items. EMO, SOC, COG, and BOD (PDQ-39se), and EN, PA, SI, and NHPD (NHP) exhibited proof of unidimensionality. There were a total of eight misfitting items in the PDQ-39se and three in the NHP. Overall, items displayed an interpretable hierarchical pattern. As opposed to the PDQ-39se scales, whose ranges of measurement are made up of the response category step measures, the dichotomous NHP scales yield only a single measure per item, thus covered a generally less wide range of measurement than PDQ-39se scales (Fig. 5), although the difference in several cases was subtle. Distances between PDQ-39se and NHP item measures (Fig. 5) and PDQ-39se item response category measures were not equal. Plots of PDQ-39se and NHP scale raw scores against their logit measures typically resulted in ogive-shaped curves of various slopes, indicating lack of scale raw score linearity.

Examination of response category step measures revealed some disordering in all but one PDQ-39se scale (EMO). Disordering of response category step 1 was observed in all but two scales, being most common in the ADL (three items) and COG (two items) scales. Step 2 was disordered in four items of the MOB and one of the COG scale. Items 28 and 29 of the SOC scale exhibited disordering steps 3 and 4, and step 4, respectively. Response probability curves of PDQ-39se items revealed that response categories were missing or co-modal with adjacent categories in 26 of the 39 items. These typically involved categories 1 and/or 2 (18 instances), but also categories 3 and/or 4 (eight instances).

PDQ-39se scales were able to separate measured respondents into a larger number of statistically distinct groups than NHP scales. All PDQ-39se scales but one (SOC) exceeded two person strata, and three (MOB, ADL, and EMO) exceeded the clinically recommended threshold of three person strata. The latter was true also for the NHPD, whereas all but one other NHP scale (ER) failed to exceed two person strata. Mean person logit measures indicated that all scales but one (EN) measured at a level corresponding with more impaired health than that of the performance of the measured sample. Although within one SD, the PDQ-39se tended to show a more pronounced dislocation in targeting its measured respondents than the NHP did.
Fig. 5 Item logit measures, measurement range, hierarchical item order, and unidimensionality of (a) PDQ-39se and (b) NHP scales according to Rasch analyses. Numbers correspond to item numbers of the full PDQ-39se and NHP, respectively (paper VI). Arrows indicate the highest and lowest item response category step calibration for PDQ-39se scales (a). Items representing more impaired health are to the right of items representing less impaired health status. Circled numbers represent items that do not fit the unidimensional measurement model. Item 34 of the NHP was not endorsed by any respondents. Data from paper VI.
All PDQ-39se scales displayed signs of potential DIF between older and younger, as well as male and female, respondents. Two NHP scales did not show any signs of DIF between genders (EN and SI) or age groups (SL and EN). Overall, 35% of PDQ-39se item threshold measures and 24% of NHP measures displayed signs of gender DIF. Corresponding figures for age were 41% and 14%, respectively. The NHPD showed signs of DIF for both age and gender in 13% of items.
One of the first symptoms of an approaching nervous breakdown is the belief that one’s own work is terribly important

Bertrand Russell
5. DISCUSSION

This work has addressed the assessment of DA-ergic cell replacement therapy for PD at three different levels by the development, evaluation, and application of assessment methods aimed at four different aspects. At the level of body structure and function, synaptic DA-release from grafts implanted into the putamen has been monitored by application of a novel radiolabel brain imaging technique. A set of traditional means of clinical quantification of parkinsonian symptomatology has been applied, and a novel clinical rating scale for dyskinesias in PD has been developed, evaluated, and applied at the level of impairment. The use of a generic health status questionnaire, tapping various constructs predominantly at the level of activity limitations and participation restrictions, has been explored in grafted PD patients, and this questionnaire has also been evaluated and compared with a PD-specific health status questionnaire regarding feasibility and psychometric properties. The findings illustrate the value and importance of a multi-level approach to the assessment of cell replacement therapies, and are in general support for the usefulness of the employed means of assessment. However, most of the methods used here will require further development, refinement and/or evaluation before they can be considered optimal, and additional areas of outcome not tapped thus far will need to be addressed in future studies.

The following sections will discuss specific findings of the current study, as divided into the various areas of assessment: motor function, graft function, dyskinesias, and perceived health. This will be followed by a more general discussion addressing some specific issues related to outcomes measurement in restorative interventional approaches to PD, various implications and needs for further work with regard to the areas focused on in this study, as well as taking a look beyond the current paradigm.

5.1. Clinical assessment of motor function

Clinical and PET scan observations in five parkinsonian patients (paper I) indicate that sequential transplantation in PD does not compromise survival of either the first or the second graft. Additional improvements were evident in three out of five patients. The two other patients, who had shown a poor response also to the first graft, developed atypical features during the follow-up period. The magnitude of symptomatic relief exerted by the second graft was less pronounced than after the first one but followed the same pattern, with patients exhibiting bilateral improvements that were more pronounced contralaterally. These findings show that bilateral grafts give rise to more pronounced symptomatic relief as compared to unilateral ones. However, the improvement is still incomplete and varies among patients.
Early clinical transplantation trials were characterized by diversity in terms of their clinical assessment protocols (Lindvall 1994). Such practice prevents virtually any valid comparisons of results from different centers employing various assessment procedures (Diamond and Markham 1983). For these reasons, the CAPIT protocol was agreed upon and has been used, in whole or in part, in the vast majority of clinical transplantation trials since its publication in 1992 (Langston et al. 1992).

The core CAPIT components of the clinical assessment of motor function are the UPDRS and four timed motor tests. Although several issues remain in doubt or are in need of refinement (see below, section 5.5.1.), the UPDRS has been evaluated regarding various aspects of its psychometric properties (Goetz et al. 1995; Martínez-Martín et al. 1994; Richards et al. 1994; Siderowf et al. 2002; Stebbins et al. 1998, 1999). In contrast to the hand/arm movement between two points and the walking test (see above, section 1.2.2.2.), formal evaluations of the finger dexterity and pronation/supination tests are sparse and the relevance of both has been questioned (Lang et al. 1995). The main concern is that they do not necessarily reflect the level of bradykinesia, because the time taken to complete them is influenced not only by the speed but also by the quality of movements. The CAPIT protocol has recently been revised with the aim to make it applicable to all neurosurgical interventions in PD, not just transplantation trials. The revised version is called the Core Assessment Program for Surgical Intervenional Therapies in Parkinson’s Disease (CAPSIT-PD; Defer et al. 1999). In response to previously expressed criticism, the pronation/supination and finger dexterity tests were omitted in CAPSIT-PD. However, both we (Wenning et al. 1997) and others (Freeman et al. 1995; Hauser et al. 1999) have used the pronation/supination test as the quantitative measure of bradykinesia in the assessment of graft function. According to these empirical experiences, it has been a useful parameter. The face validity of this test was evident also in the present study (paper I), in that changes were detected following the first, as well as the second graft, and these changes appeared similar, although occurring with various magnitudes. Thus, improvements were evident bilaterally but were more pronounced contralaterally to the grafts following both implantations. Furthermore, in patients in whom neurophysiological measures of movement time were performed, the outcome of both modalities of assessment, as well as of clinical ratings according to the UPDRS motor score, appeared similar (paper I) and the correlation between the two former was good in two and moderate in one of these patients. These observations are not sufficient to draw any firm conclusions regarding the validity of the pronation/supination test. Indeed, assessment of the responsiveness of various outcome measures, based on data from another group of patients (Table 7; Brundin et al. 2000), indicates that this motor test is one of the least responsive, with a moderate ES of about 0.5, which may be seen as an indication that this test does not constitute an optional measure of bradykinesia.
5.2. Assessment of graft function

In a long-term follow-up study of one of the patients that has exhibited the most dramatic symptomatic relief following transplantation, it was demonstrated that DA release from intrastriatal grafts of human embryonic mesencephalic tissue can be visualized and measured in vivo in patients with PD, by use of non-invasive imaging of D2 receptor occupancy in the presence and absence of a DA-release stimulating drug challenge. These findings indicate that well-developed nigral grafts can restore both basal and drug-induced DA release to normal levels.

At ten years after transplantation, FD uptake was restored to normal levels in the grafted putamen compared to a level of 12% of normal on the non-grafted side. The grafted putamen had normal RAC binding, whereas D2 site availability was upregulated in the non-grafted putamen. These data indicate that the new transplant-derived innervation tonically releases sufficient DA to restore D2 receptor occupancy to normal levels in the previously denervated striatum. In untreated PD patients, PET and post-mortem autoradiographic studies have shown upregulation of DA D2 receptor binding in the putamen (Sawle et al. 1993; Seeman et al. 1987), which reverses after L-dopa treatment (Antonini et al. 1994). The elevated D2 receptor binding in the non-grafted putamen of this patient was not reversed by his recent low-dose L-dopa treatment. The normal level of D2 receptor binding in the grafted putamen is consistent with studies after transplantation in animal models of PD, where DA-ergic grafts that reinnervate the striatum normalize upregulated D2 receptor binding (Brundin et al. 1994; Elsworth et al. 1998), most probably due to continuous stimulation of postsynaptic sites by DA released from the graft-derived axon terminals. Indeed, in vivo monitoring of DA release in the rat PD model has demonstrated that grafted nigral neurons can restore baseline striatal DA levels to control values (Brundin et al. 1988). Microdialysis studies in non-human primates indicate that a 25-30% decrease in striatal D2 receptor binding (measured by SPECT) following amphetamine administration reflects a 2- to 10-fold peak increase in extracellular DA levels (Laurelle et al. 1997). Therefore, the amphetamine induced 26.6% reduction in RAC binding detected in the grafted putamen of this patient is likely to reflect a similar level of increase in extracellular DA derived from the transplanted neurons. The small, 4.5% reduction of RAC binding in the non-grafted putamen confirms the low capacity for DA release in this severely denervated structure, and is similar to that seen in non-grafted PD patients (Piccini et al. 2000a).

Taken together, our observations in this patient suggest that in the best cases, nigral transplants reinnervate the putamen extensively, to a level sufficient to normalize both FD uptake and D2 receptor occupancy throughout the grafted putamen. It seems highly likely that the efficient restoration of DA neurotransmission in large parts of the grafted putamen underlies the patient’s major and sustained symptomatic improvement, reflected by, e.g., about a 50% decreased UPDRS motor score while off medications.
However, the exact extent to which graft-derived DA release influences the magnitude and pattern of clinical response remains to be clarified. This study also provides the first evidence that well-developed and clinically efficacious nigral transplants restore both basal and drug-induced DA release in the striatum of PD patients. Our data thus support the usefulness of the RAC/PET method for in vivo analysis of DA release from grafted DA-producing cells after transplantation. This modality of assessment of graft function is highly needed because, in addition to embryonic mesencephalic tissue, other sources of DA-producing cells, including, e.g., adrenal medulla, sympathetic ganglia, carotid body, embryonic xenogeneic porcine mesencephalic tissue, genetically engineered fibroblasts, and progenitor and stem cells, have been considered potentially useful for grafting in patients (for refs and review, see Brundin and Hagell 2001; Lindvall and Hagell, in press). The RAC/PET method is a novel powerful tool in the further development of cell replacement strategies in PD.

5.3. Clinical assessment of dyskinesias

5.3.1. Development and evaluation of the Clinical Dyskinesia Rating Scale

The CDRS was devised in response to the lack of an established means of systematic assessment of dyskinesias at an appropriate level (i.e., impairment), incorporating separate ratings of the major types of dyskinesias (hyperkinesias and dystonia) and allowing for evaluation of their topographical distribution, while avoiding additional patient burden. The resulting scale, which uses a non-defined 5-stage scoring code, was evaluated for inter- and intrarater reliability among clinicians experienced with PD. Interrater reliability was also assessed for an alternative form of the scale, using a defined scoring code. Ratings performed with the two scoring codes did not differ either for hyperkinesias or dystonia ratings, and reliability indices support that CDRS yields reliable ratings of dystonia within, and of hyperkinesias within as well as between raters. However, discrepancies were observed between raters’ dystonia assessments. This observation is in agreement with results from the evaluation of the Rush dyskinesia scale, which ascribes identification of the types of dyskinesias observed (defined as “chorea”, thus largely comprising hyperkinesias in the CDRS, “dystonia”, or “other”), as well as the most disabling type (Goetz et al. 1994; see above, section 1.2.2.4.). These authors found that intrarater agreement was acceptable (Cramér’s V: 0.836 - 0.840) or suboptimal (0.699 - 0.732) for classification of dyskinesias present and identification of the most disabling type of dyskinesia, respectively. Interrater agreement, however, was low for both these judgements (κ: 0.279 - 0.483).

In addition to the evaluations performed in the original analysis of the CDRS (paper III), data from six raters were re-assessed regarding reliability of hyperkinesia and dystonia ratings using the generally recommended ICC. The results from these analyses
are in general agreement with those initially obtained by means of Kendall’s $T$ and $W$, thus confirming acceptable levels of inter- and intrarater agreement for ratings of hyperkinesias, as well as of intrarater agreement for dystonia ratings. However, the disagreement between raters on dystonia scores is even more pronounced according to the ICC (Table 6). The relationship between these ratings and clinicians’ global impression of the overall levels of dyskinesia and parkinsonism exhibited by the patients during each of the rated video sequences was also explored (Table 6). Strong positive and negative correlations between hyperkinesia scores and the levels of dyskinesia and parkinsonism, respectively, indicate that hyperkinesias are associated with low levels of parkinsonism (hence, the “on” phase) and support the validity of this part of the scale. Moderate correlations were seen between dystonia scores and the global levels of both dyskinesia and parkinsonism, being somewhat stronger for parkinsonism. This pattern is also conceivable because although dystonia typically appears during the “off” phase, dystonic features are not uncommon during “on” (Durif et al. 1999; Luquin et al. 1992; Quinn 1998; Vidailhet et al. 1999). The use of a global CDRS score to provide an overall estimate of the severity of dyskinesias was also evaluated regarding reliability and relationship with the raters’ global impressions (Table 6). Rater agreement was found acceptable and its correlational pattern supports its validity.

There are several possible explanations to the inconsistent rating and classification of dystonia among clinicians observed both by us and others (Goetz et al. 1994; paper III). In both studies, ratings were made from videotapes. There are several shortcomings associated with video assessments, but also advantages. For example, it allows for blinded ratings and retrospective assessments of available video recordings. However, it is possible that a better distinction between various types of dyskinesias is achieved from clinical “live”-ratings. Secondly, these observations may be due to the lack of clear rater instructions, since this is provided neither by the CDRS nor the Rush scale (Goetz et al. 1994; paper III). However, our attempt to clarify the dystonia section of the CDRS and provide definitions of the degree of interference of movements, as modified from the Rush dyskinesia scale (Goetz et al. 1994), did not yield any apparent advantages (paper III). The defined version of the CDRS yielded a marginally higher level of concordance than the non-defined one for ratings of dystonia but not for hyperkinesias. In general, it should be advantageous to have clearly expressed guiding definitions anchored to the respective scores in a scale (Lang 1995). One reason for the observed lack thereof may be that the definitions used for the second evaluation of the CDRS (Table 5) are inadequate and do not provide the guidance necessary to improve consistency among raters. Indeed, several of the raters who used the defined scoring code commented that it was unclear how to use it in practice and one of the raters, who also had used the non-defined version, thought that ratings were more difficult with the definitions (unpublished observations). While the score definitions refer to dyskinesia interference with voluntary movements, there is no recognition of the fact that none of the tasks performed by the patients while being assessed involves
either the neck or the trunk. Furthermore, while motor activation indeed typically induces or aggravates dyskinesias, this worsening is often seen in body parts other than the activated one (Durif et al. 1999; Vidailhet et al. 1999). Again, the evaluated scoring definitions do not take this into consideration but refers only to the activated body part. Finally, the lack of concordance may illustrate disagreement among clinicians regarding the definition of dystonia, or specific difficulties related to rating severity of dystonia. Taken together, in the choice between the two presently evaluated scoring codes, these observations and experiences speak in favor of using the non-defined scoring code. However, there is a need for identification of clear and practically useful scoring definitions in order to facilitate ratings and promote rater consistency, for example by introducing dystonia- and hyperkinesia-specific score definitions to be used in the respective sections of the scale. Until these issues have been resolved, rater reliability should be established and documented whenever using the CDRS in clinical research.

5.3.2. Application and practical aspects of the Clinical Dyskinesia Rating Scale

In a subsequent application of the CDRS, the influence of intrastriatal grafts of DA-rich human embryonic mesencephalic tissue on dyskinesias was retrospectively assessed from available video recordings in 14 PD patients, preoperatively and up to 11 years after grafting (paper IV). Dyskinesias were found to increase significantly during the “off” phase after transplantation. The severity of dyskinesias was not related to the magnitude of graft-derived DA-ergic reinnervation or symptomatic relief, but to the amount of tissue implanted into the putamen. Dyskinesias also tended to correlate with the degree of striatal DA-ergic denervation preoperatively, and to the number of putamen trajectories. Two patients implanted with tissue that had been stored for up to 8 days before surgery developed the most pronounced postoperative dyskinesias. This study shows that present grafting procedures can induce “off” phase dyskinesias. This adverse effect is probably dependent on postsynaptic striatal alterations due to the preoperative loss of DA-ergic input, non-DA-ergic components within the grafted tissue, and/or the intrastriatal surgical trauma, but may also be caused by small grafts giving rise to patchy reinnervation. These observations argue against the notion that “off” dyskinesias are a characteristic feature of DA cell replacement per se, and provide no evidence that this side effect should stop the further development of a cell replacement therapy for PD. However, the underlying mechanism(s) must be better understood so that “off” dyskinesias following neural transplantation can be avoided.

In addition, the results from this study allow for consideration of some further aspects of the CDRS and its use. First, as advocated by several authors (Lang et al. 1995; Nutt 1999; Vidailhet et al. 1999) separate ratings of various anatomical regions provide clear advantages over single, overall scores. It has previously been demonstrated that
the single-score approach is under-powered in detecting altered levels of dyskinesias. Uitti et al. (1997) thus compared the Rush scale and the Mayo dyskinesia rating scale (a modification of AIMS including separate ratings of different body parts) in an evaluation of unilateral pallidotomy for PD. The Rush scale was unable to detect any changes, whereas the Mayo scale showed clear and significant postoperative reductions of dyskinesias. This advantage of separate anatomical ratings was also evident in this study (paper IV), as illustrated by significant alterations in scores of hyperkinesias and dystonia, as well as the overall level of dyskinesias (Fig. 4). In addition, such ratings also allow for analyses of the topographical distribution of dyskinesias (Marconi et al. 1994; Vidailhet et al. 1999). Although not analyzed quantitatively in this study (paper IV), a clear description of the distribution of dyskinesias was enabled. Second, there is a clear advantage to perform separate ratings of the two main types of dyskinesias, defined as hyperkinesias (predominantly representing choreiform movements) and dystonia in the CDRS. Although not reaching statistical significance (due to the relatively small sample size), differential developments of dystonia and hyperkinesias could be observed over time (Fig. 4). Indeed, several other reports have also indicated particular effects on dystonia and hyperkinesias following therapeutic interventions (Carpentier et al. 1996; Limousin et al. 1995, 1996; Quinn and Marsden 1986). Furthermore, indices of neurobiological differences between these two main types of dyskinesias have also been evident in that terminal activity in the motor component of the subthalamic nucleus (STN) was found increased in dystonic but not in choreic L-dopa or apomorphine treated monkeys with MPTP-induced parkinsonism, whereas the activity in the non-motor portion of the STN was increased in both subsets of monkeys (Mitchell et al. 1992).

Taken together, these studies support the validity and appropriateness of the approach to dyskinesia assessment advanced by the CDRS. Despite of the current problem with discrepancies between raters’ dystonia scores (see above, section 5.3.1.), the CDRS constitutes the most optimal approach to its intended modality of dyskinesia assessment, fulfilling clinical requirements that no other psychometrically evaluated dyskinesia rating scale does. Indeed, this type of a dyskinesia rating scale has been warranted for some time (Melamed et al. 1999; Nutt 1999, Rascol 1999) and is currently recommended for assessment of dyskinesias following neurosurgical interventions for PD (Defer et al. 1999). It should also be underscored that the usefulness of the CDRS is restricted to assessment of the amount of dyskinesias at a given time. For more comprehensive assessments, information regarding the daily duration and the activity limitations/participation restrictions related to dyskinesias needs to be obtained by other means.
5.4. Assessment of perceived health

5.4.1. Perceived health in grafted PD patients

Explorative use of the NHP (including the NHPD) in five PD patients before and up to 24 months following bilateral intrastrital implantation of human embryonic DA-rich mesencephalic tissue revealed compromised preoperative and improved postoperative perceived health and/or distress in all cases. This, together with the postoperatively observed alternations, allow for two main conclusions. First, intrastrital grafts of embryonic DA-ergic tissue are able to induce improvements within a variety of aspects of perceived health, as assessed using the NHP. In this group of patients, preoperative impairments were seen within all dimensions of the NHP, although not all patients were impaired on all dimensions. Postoperative follow-up demonstrated a statistically significant improvement within one dimension (PM). Although failing to reach statistical significance, due to the small sample size, the other dimensions of the NHP, as well as the NHPD, also showed clear postoperative alterations. A possible pattern of perceived health outcomes, consisting of improved mobility and emotions, along with decreased distress and increased energy could be identified following transplantation. Second, these data support the notion that health status questionnaires provide valuable information on the outcome of intrastrital transplantation in PD additional to that obtained by traditional study protocols such as the CAPIT. Such questionnaires address important aspects of PD, including non-motor features such as emotional well-being, fatigue, and sleep disturbance, that are not captured by means of observer-derived clinical neurological assessments or ratings.

Re-assessment of the responsiveness of the various NHP scales (table 7) revealed that four of the scales (ER, EN, PM, and NHPD) displayed large ESs, whereas they were moderate in two (SL and PA) and small in one scale (SI). The most responsive dimension of the NHP was that covering physical mobility (PM). Similarly, the UPDRS motor score was the observer-derived measure that exhibited the largest ES. These observations are not surprising, because it can be presumed that striatal DA replacement primarily should improve patients’ mobility. However, it is interesting to note that of these two measures of overall mobility or motor function, the patient-derived PM scale of the NHP showed better responsiveness than the observer-derived motor examination of the UPDRS (Table 7). Examination of the other ESs provide further interesting information. The large ESs associated with emotional and fatigue related measures (ER and EN, respectively), as well as with that of overall illness-related distress (NHPD), indicate that patients improved to a large extent within these areas following transplantation. Data from around six and twelve months after grafting indicate that these improvements, in general, were progressive, thus broadly following that of striatal graft maturation, as assessed using FD PET (Brundin et al. 2000; paper V). Hypothetically, these observations could thus indicate that these non-motor problems,
directly or indirectly, may be manifestations of the striatal DA deficiency in PD. This renders further studies on the relationships between variables of perceived health and the biological pattern of disease-related denervation on one hand, and the intervention-induced pattern of reinnervation on the other, highly warranted in order to gain a better understanding of these common consequences of PD. However, in order to allow for valid studies like that, the psychometric properties and dimensionalities of health status questionnaires must be ascertained.

5.4.2. Evaluation of generic and disease-specific health status questionnaires

The feasibility and psychometric properties of a generic (NHP) and disease-specific (PDQ-39) health status questionnaire were evaluated and compared from a pilot postal survey. In addition, the linguistic and content validity of the Swedish language version of the PDQ-39 (PDQ-39se) were addressed in a small clinical sample. The clinical study indicates that the PDQ-39se exhibits an acceptable, albeit sub-optimal, level of content validity as a measure of health, functioning, and well-being in PD. Although a majority of the PDQ-39se may appear linguistically valid, with only four out of 42 components (39 items + instructions, frame question, and response alternatives) showing substantial problems, important shortcomings were obvious. These are of particular importance because two of the problem areas (the frame question and the response alternatives) may influence all other items in the questionnaire. The other two (items 28 and 29, both belonging to the SOC scale) can give rise to unintentionally inverted responses. These observations have potential implications also for other language versions of the PDQ-39 since these only rarely (Bushnell and Martin 1999) appear to have been evaluated regarding linguistic validity for its target population, and underscore the importance of documenting both the linguistic and psychometric qualities of new language adaptations prior to their use in clinical research or trials (Bullinger et al. 1993, 1998; Fayers and Machin 2000; Hunt et al. 1991; Leplége and Verdier 1995; McKenna and Whalley 1997).

The observed psychometric properties of the PDQ-39se were similar to those reported for the British source version and other translations (Bushnell and Martin 1999; Damiano et al. 2000; Jenkinson et al. 1995, 1998; Katsarou et al 2001; Martínez-Martín et al. 1998, 1999b; Peto et al. 1995; Peto and Jenkinson 1999). However, in several instances (e.g., coefficient $\alpha$ and known-groups validity) its performance does not meet recommended standards. In addition, we also addressed a number of areas rarely or never reported before. Various PDQ-39se scales thus exhibited floor effects and suboptimal ability to separate subjects into statistically distinct groups, as well as a measurement bias toward the more severe end of the health continuum. Non-additivity of item scores and non-linearity of scale scores, along with questionable dimensionality, response category ambiguities, and signs of possible DIF between gender- and age...
groups, were also observed. Except for suggested DIF, the present findings support the validity of the EMO scale, whereas indications of additional shortcomings were obtained for all other PDQ-39se scales, in particular the SOC scale, which failed to fulfill virtually any criteria. The generic NHP displayed similar results but was associated with more floor effects and poorer ability to separate subjects into statistically distinct groups than the PDQ-39se. However, measurement bias toward the more severe end of the health continuum among subjects with non-extreme scores was less pronounced for NHP scales, and there were fewer indications of suspect dimensionality than in the PDQ-39se scales. Data support the validity of the NHPD, which, except for some suggested DIF, performed adequately according to all predefined criteria.

These observations carry important messages in terms of the usefulness of these questionnaire scales as outcomes measures. The questionable unidimensionality and construct validity of four PDQ-39se and three NHP scales challenges their interpretability. If scales are not unidimensional and fail to achieve construct validity it is unclear what a certain score means or represents. This limitation is of particular relevance if attempting to use such scales as outcome measures in emerging therapeutic principles, such as cell replacement therapy in PD, aimed at restricted areas and/or neuronal populations within the brain, or in studies of the relationship between non-motor aspects of PD and patterns of denervation. The finding that neither the PDQ-39se nor the NHP appear suitable for use in the earlier stages of the disease, as indicated by large floor effects in mildly perceived PD, and their bias towards the more severe end of health impairments are also important. This is of particular concern in perspective of their usefulness in, e.g., neuroprotective therapeutic trials, where it is likely that the primary target will be patients in relatively early stages of the disease. In this scenario, our data do not suggest that either the NHP or the PDQ-39 will be able to provide the necessary means to monitor disease progression from the patient’s perspective or in terms of important non-motor PD problems (Larsen et al. 2000; Schulman et al. 2002) that are addressed by these questionnaires (e.g., sleep disturbance, fatigue, and feelings of depression).

Taken together, these observations illustrate several important problems related to measurement validity and underscore the necessity of thorough and systematic evaluations of health status questionnaires, as well as other outcome measures. Relative strengths as well as weaknesses are apparent for both the PDQ-39se and NHP, and it is difficult at this time to conclude that one is superior to the other for use in PD. Although both questionnaires bear promise as disease-specific and generic health status questionnaires in PD, respectively, further developmental work and evaluations are needed. Thus, our observations indicate that the dimensionalities of these instruments need further consideration. It may, for example, be possible to obtain more well-functioning and interpretable scores by reducing the number of scales and items in the questionnaires. This will require further analyses on larger samples. No information is
yet available about the test-retest reliability of the PDQ-39se, with good reproducibility being essential for a questionnaire to be considered useful. This is true also for the NHP when used in PD, although it has been evaluated regarding this aspect in other patient groups (Anderson et al. 1993; McDowell and Newell 1996; Wiklund 1989). As a result of the current observations, the PDQ-39se has undergone revision. Before sufficient evidence in support of the psychometric properties, including test-retest reliability, of the revised PDQ-39se has been documented, neither version can be considered valid as an inferential outcome measure.

5.5. General discussion

5.5.1. Some selected issues in assessing restorative approaches to Parkinson’s disease

While this study has taken a multi-level approach to the assessment and outcome measurement of cell replacement therapy in PD, the specific areas of focus have been restricted and a number of aspects have been left unattended. These involve, for example, graft function beyond that of DA storage and release, more specific evaluations of the validity and reliability of timed motor tasks such as the pronation/supination test, evaluation of the UPDRS, in particular the commonly used motor examination section, additional refinement and evaluation of the CDRS, and application and evaluation of objective measurement of activity limitations. In addition, frequent non-motor problems such as cognitive decline, depression, and autonomic dysfunction, have not been addressed in this work.

To broaden our understanding of the requirements for a successful cell replacement therapy for PD, it is important to consider graft function beyond that of DA storage and release. We thus recently addressed the issue of functional graft integration within the host brain in PD patients (Piccini et al. 2000b). Movement-related activation of two frontal cortical areas, the supplementary motor area (SMA) and the dorso-lateral prefrontal cortex (DLPFC), was studied using regional cerebral blood flow measurement and PET in four patients grafted bilaterally in the caudate and putamen. The SMA and DLPFC are known to be important in the preparation and selection of voluntary movements, their function is influenced by the basal ganglia-thalamo-cortical neural circuitries, and their impaired activation is believed to underlie bradykinesia. Preoperatively, there was only a small activation of the SMA and no significant activation of the DLPFC. No significant differences in activation were observed 6.5 months after grafting as compared to preoperatively, while at 18.3 months there was significantly increased activation of both. The time course of clinical improvement paralleled that of the increase of cortical activation with partial recovery after 6.5 months and substantial improvement at 18.3 months. In contrast, striatal FD uptake had increased already at 6.5 months and showed no further change at 18.3 months.
after grafting. These findings, together with those in paper II, indicate that successful grafts in patients with PD improve striatal DA-ergic neurotransmission and restore movement-related cortical activation, which probably is necessary to induce substantial clinical improvement. These data also provide new evidence that the functional effects of the grafted neurons go beyond those of a simple DA delivery system. Restoration of non-regulated DA release, as in the early stages of graft maturation, when FD uptake is already significantly elevated, seems to be insufficient to improve cortical activation during movement and to induce substantial clinical recovery. In order to increase basal ganglia-thalamo-cortical neurotransmission and movement-related cortical activation, the grafted DA neurons probably need to establish both efferent and afferent synaptic connections with the host brain.

Clinical assessment of various aspects of PD is a delicate task because the disease presents with a number of complicated aspects that introduce challenges rarely seen in other disorders. First, the symptomatology of PD is vast. The list of symptoms and signs encountered in PD can be made virtually endless and relatively little is known about their underlying pathophysiological substrates (Jankovic 1992; Olanow et al. 2001). Although many clinical features are presumed to be direct or indirect consequences of the nigrostriatal degeneration and/or the cardinal PD symptoms, these relationships are thus far only understood in part, and the neurobiology of PD goes far beyond nigral DA cell degeneration and striatal DA deficiency (Lang and Lozano 1998a). Second, the symptomatic profile differs between patients and although rough classifications can be made, e.g., tremor- vs postural instability and gait difficulty (PIGD) types (Jankovic et al. 1990), in practice virtually any given patient presents a unique set of symptoms, signs, and disease-related problems. Third, it is not uncommon that patients experience periods of improvement or worsening, lasting for a few days or weeks (Quinn 1998). Fourth, after some years of treatment many patients experience short-term daily drug response fluctuations with disabling “off” periods and dyskinesias (see above, section 1.1.7.). Taken together, these aspects introduce specific demands regarding clinical assessment of disease presentation, progression, and response to therapeutic interventions. One situation where these challenges are particularly evident is when attempting to understand and develop a therapeutic principle that aims at restoring the underlying pathophysiological deficit by focal intracerebral replacement of lost cells. First, it is unclear if, or to what extent, intrastriatal DA-ergic cell replacement may influence neurotransmitter alterations beyond that of the striatal DA deficiency. Second, because the relationships between various clinical features and brain abnormalities are poorly understood, the possible outcomes of cell replacement need to be considered at a broad level using methods that are sensitive and specific in order to determine the therapeutic potential, as well as gaining insight into the neurobiology of the disease. Thus, clinical assessment of specific graft-induced effects in PD patients should ideally tap a large spectrum of specific consequences of the disease with sufficient accuracy, as well as providing information on the overall
outcomes, while still being feasible and not introduce unnecessary patient and investigator burden. In order to fulfill these needs, the methods used should be feasible, well defined, reliable, and valid.

Observer-derived clinical rating scales are sometimes referred to as “objective” (e.g., Goetz et al. 1994). However, this is not a valid assumption. Although the clinician must be considered a better and more valid assessor of impairments, because all clinical symptoms and signs are not fully appreciated by the patient (Brown et al. 1989; Golbe and Pae 1988; Hagell and Sandlund 2000), this does not render such observations more objective than, e.g., self-assessments of participation restrictions. Rather, the subjectivity has been shifted from the patient to the clinician. Indeed, objective measurement has recently been defined by the Program Committee of the Institute for Objective Measurement (2001) as: “the repetition of a unit amount that maintains its size, within an allowable range of error, no matter which instrument is used and no matter who or what relevant person or thing is measured”. One practical approach to enable objective measurement is the Rasch measurement approach (Bond and Fox 2001). As partly illustrated in paper VI, this method enables evaluation and construction of linear variables that are unidimensional and freed from the influences of the measured subjects and scale items. These are fundamental features of objective measurement. Unidimensionality, for example, is necessary in order to measure and understand treatment outcomes, regardless whether they are perceived or observed (see also above, section 5.4.2.). Within the area of rehabilitation, Rasch methodology has been successfully employed in the development of observer-derived objective measures of activities of daily living (e.g., Fischer 1993; Linacre et al. 1994), as well as in the application of measures of neurological impairment (Roth et al. 1998). Given the potential of this measurement approach and the need for measures of high quality and specificity, in particular when assessing the outcomes of focal intracerebral interventions or attempting to understand relationships between pathological processes monitored in vivo and clinical manifestations, it would indeed be warranted to apply this technique in order to devise refined and objective measures for use in clinical PD research and therapeutic trials. The problems associated with currently available rating scales can be illustrated by the UPDRS. Although empirical experience support the general usefulness and validity of the UPDRS (see above, section 1.2.2.2.), further development and refinement of this scale are needed, in particular regarding its dimensionality and item selection (van Hilten et al. 1994). The main objections are that it is too lengthy, and, in part, redundant (Martínez-Martín et al. 1994; van Hilten et al. 1994). For example, in one study (van Hilten et al. 1994) it was shown that both sections II and III of the scale could be substantially abbreviated (by about 50%) without loss of reliability or validity. Furthermore, recent factor analyses of the motor examination section (part III) of the UPDRS have displayed clear signs of multidimensionality (Stebbins et al. 1998, 1999). While it, in general, must be considered a valid tool for assessment of parkinsonian symptoms and signs, it is still unclear how best to use the UPDRS in
various situations. Given past experiences from other clinical areas (e.g., Fisher 1993; Linacre et al. 1994; Roth et al. 1998), it is reasonable to assume that these shortcomings could be circumvented by thorough developmental work using, e.g., the Rasch measurement approach. Not only would this yield unidimensional and linear measures that allow for more accurate interpretation and understanding of patterns of disability and recovery (Granger and Linn 2000) and use of more powerful data analytic approaches (Merbitz et al. 1989; Wright and Linacre 1989), but it would most probably also result in more brief and effective measurement tools (Linn et al. 1999).

These concerns are not restricted to the UPDRS and self-administered questionnaires of perceived health, but apply also to other means of assessment, such as the CDRS. Before determining the dimensionality and measurement properties of the CDRS, there is, however, one main issue that needs to be resolved, i.e., the inconsistency of dystonia ratings, probably by defining specific scoring codes and instructions for the two respective sections of the scale (see above, section 5.3.1.).

5.5.2. Beyond the current paradigm

It is conceivable that the consequences of disease and influences of therapeutic intervention go beyond that of the current paradigm. There is thus a need for assessment and, hence, conceptualization, of an overall level of impact of disease from the patient’s perspective. Such a level may or may not represent QoL. One approach that repeatedly and successfully has been employed over the past decade is the development of outcomes measures according the needs-based model of QoL, which goes beyond the frameworks of the ICIDH and ICF (McKenna et al. 2000a). This model grew out of work with patients suffering from depression (Hunt and McKenna 1992) and has subsequently been applied successfully in the development of disease-specific instruments for a wide range of diseases (McKenna et al. 2000a, 2000b). The model argues that life gains its quality from the ability and capacity of individuals to satisfy their needs. Functions may well be influential but only where they provide the means by which needs are fulfilled. The model argues that QoL is good when most needs are met and poor when few needs are being satisfied. In support of this approach is that measures devised for various diseases have demonstrated unidimensionality, a prerequisite also for overall outcome measures.

An extension beyond the affected person will involve family members and others close to the patient, as well as the society at large. Although these areas have been largely unattended in the past, important work on the burdens of caregivers to PD patients has now been initiated (Carter 1999) and initial studies on the impact of PD on the patient, family, and society have presented data showing a considerable burden for all three parties (Whetten-Goldstein et al. 1997). Societal consequences of disease...
primarily involve the burden posed by health care costs, loss of productivity, and resource use adjunct to that of the healthy population. Given the current development of escalating health care expenditures, shrinking resources, and rising life expectancy (Anell and Willis 2000; Wise 1998), this aspect of outcomes research is likely to become increasingly important as new and potentially costly treatments, such as cell replacement therapy for PD, are developed. The economic impact of therapeutic interventions for PD has so far been sparsely reported in the literature. However, evaluations from the US indicate higher costs, but also better effectiveness, of the DA agonist pramipexole, as compared to baseline treatment in early as well as advanced PD, whereas pallidotomy would reach the same cost-effectiveness only if procedure costs were reduced by two-thirds or if the postoperative utility was equivalent to being restored to normal health (Hoerger et al. 1998; Siderowf et al. 1998). In another American-based study, the cost effectiveness of DBS of the globus pallidus and STN, as compared to best medical management, was found uncertain unless providing at least 30% clinical improvement (Tomaszewski and Holloway 2001). In order to evaluate the economic implications of new interventions in an informed manner, it is important to consider the current economic impact of the disease. Health economic studies on the cost of PD have begun to emerge during recent years (Dodel et al. 1998; LePen et al. 1999; Whetten-Goldstein et al. 1997). From a Swedish perspective, we have recently estimated the mean annual societal cost of PD to SEK 124,000 per patient, of which indirect costs accounted for 42%, followed by home care and direct health care costs, accounting for 34% and 23%, respectively. The dominating health care cost was drug therapy, which accounted for 44% of the direct medical costs and 10% of the total costs (Hagell et al., submitted for publication). These findings confirm that PD causes a considerable societal burden and provide valuable baseline data for future health economic evaluations of novel interventions, such as cell replacement therapy.

5.5.3. Implications and needs for further work

This study demonstrates that when technical advances are taken advantage of it is now possible to monitor not only intrinsic and graft-derived DA cell viability, but also synaptic basic and drug-induced transmitter release from grafts in PD patients. Development, application, and evaluation of various means to assess consequences of PD and graft-induced alterations of those consequences, have provided new and valuable insights into the potentials of this therapeutic principle, but also identified a number of assessment and measurement related problems associated with these methods. Results are, however, promising and it is feasible to expect further work to be able to rectify current inconveniences. The identified problems, in combination with the high demands that need to be met by outcome measures for novel therapeutic interventions such as a cell replacement therapy for PD, underscore the need for documented, thorough and systematic evaluation of any measurement tool before it
can be considered fully valid. At a time when the requirement of evidence-based medicine is increasingly emphasized, it is of fundamental importance to emphasize also the need for evidence-based measurement, with demands on thorough and systematic evaluations according to appropriate psychometric standards. Unfortunately, there are very few outcome measures in PD that would meet such criteria today. Hopefully, the initial work presented here will continue and, by time, fulfill some of the needs for more effective, valid, and well-defined outcomes measurement in PD.
The history of science and medicine has taught us that the disappointments of today are often the prelude to tomorrow’s success

W. Maxwell Cowan & Eric R. Kandel
6. CONCLUSIONS

Based on the findings presented in this thesis, the following main conclusions can be drawn:

- Established assessment methods for evaluation of cell replacement therapies in Parkinson’s disease are useful in that they are able to detect changes in parkinsonian symptomatology. The validity and psychometric properties of these methods are, however, insufficiently documented and will require further scientific efforts.

- It is possible to measure synaptic dopamine release from intrastriatal grafts of dopamine-rich human embryonic tissue in vivo in Parkinson’s disease, and well-developed grafts can restore both basal and drug-induced dopamine release to normal levels. The exact extent to which this influences the magnitude and pattern of clinical response remains to be clarified.

- Dyskinesia assessments according to the Clinical Dyskinesia Rating Scale, with separate ratings of hyperkinesias and dystonia in various body parts, yield valuable information on detailed effects of specific interventions. However, discrepancies between raters’ dystonia ratings indicate a need for further development and evaluation, and establishment of rater consistency when using the scale.

- Intrastriatal grafts of human embryonic mesencephalic tissue can induce dyskinesias in Parkinson’s disease, but these do not appear to be linked to the extent of dopaminergic reinnervation.

- Intrastriatal dopamine-rich grafts can give rise to improvements within several motor and non-motor aspects of patients’ perceived health and distress, and such evaluations provide important information not obtained by traditional clinical assessment protocols.

- Relative strengths as well as weaknesses are associated with two widely used generic and disease-specific (NHP and PDQ-39, respectively) measures of perceived health in Parkinson’s disease, and further refinements and evaluations of both questionnaires are needed before either can be considered fully appropriate for use as inferential outcome measures.

- Multi-level assessment of cell replacement therapies is valuable and important, and the current results are in general support for the usefulness of the assessment approaches used here. However, most methods will require further development, refinement, and/or evaluation before they can be considered optimal, and additional areas of outcome not tapped thus far will need to be addressed in future studies.
Parkinsons sjukdom är en relativt vanlig kronisk neurologisk sjukdom som karaktäriseras av att nervceller i ett specifikt område djupt i hjärnan av okänd anledning successivt dör. Dessa celler producerar signalsubstansen dopamin, vilken bl.a. är nödvändig för normal rörelseförmåga. Motoriska symptom (skakningar, muskelstelhet, förlängsammade rörelser och balansstörningar) dominerar sjukdomsbilden men även andra problem såsom trötthet, sömnstörningar, depression, smärta och känslstörningar är vanliga. Medicinsk behandling med L-dopa är till en början ofta framgångsrik, men inom 5-10 år kommer flertalet patienter in i s.k. komplikationsfas. Detta innebär svängningar mellan uttalade parkinsonsymtom och normal eller ofrivillig rörlighet (s.k. dyskinesier). Sedan 1980-talet har begränsade behandlingsförsök genomförts i syfte att försöka reparera den sjukdomsdrabbade hjärnan. Detta har skett genom transplantation av omogna dopaminproducerande nervceller, tagna från elektivt aborterade mänskliga embryon, till det område i hjärnan där bristen på dopamin är som störst. Dessa studier har visat att de transplanterade cellerna överlever och växer in i värdhjärnan och att denna behandlingsprincip kan ge tydliga och långvariga förbättringar av patienternas motoriska symptom.


För tillförlitlig bedömning och utvärdering kävs att de metoder som används uppfyller vissa vetenskapliga krav. De ska således vara tillförlitliga, d.v.s. ge så exakta mått som möjligt med minsta möjliga slumpmässiga variation (reliabilitet). Vidare ska de måta det som de utger sig för att mäta (validitet) och ingenting annat (unidimensionalitet), samt ha förmåga att reflektera förändringar (responsivitet).

Målet med denna avhandling var att applicera, utveckla och evaluera metoder för utvärdering av cellterapi vid Parkinsons sjukdom inom tre olika nivåer: de transplanterade cellernas förmåga att frisätta dopamin, transplantateffekter på motoriska symptom och dyskinesier, samt effekter avseende patienternas upplevda hälsa.


Mot bakgrund av erfarenheter av befintliga kliniska skattningsskalor för dyskinesier vid Parkinsons sjukdom utvecklade vi en ny skala, Clinical Dyskinesia Rating Scale (CDRS). Enligt CDRS skattas svårighetsgraden av dyskinesier separat för sju olika kroppsdelar (fyra extremiteter, bål, nacke och ansikte) enligt en femgradig skala (0 = inga; 1 = milda; 2 = måttliga; 3 = svåra; och 4 = extrema dyskinesier). Dessa skattningar görs separat för de två huvudsakliga typerna av dyskinesier vid Parkinson’s sjukdom, hyperkinesier (ffa choreiforma, dansliknande rörelser) och dystonier (krampliknande felställningar och rörelser). Reliabiliteten testades inom och mellan bedömare (intra-respektive interbedömarskattning), som skattade 23 patientekvenser från ett videoband vid två olika tillfällen. Några av bedömarna använde även en version av CDRS där dystoniskattningen preciserades att endast omfatta dystona felställningar.

I en serie om fem patienter som transplanterades till båda sidorna av hjärnan utvärderade vi värden av att använda frågeformulär för mätning av patienternas upplevda hälsa vid cell transplantation, samt om denna behandling kan ge mätbara sådana effekter. Patienterna fyllde i det generiska hälsostatus formuläret Nottingham Health Profile (NHP) före samt upp till två år efter transplantation. Preoperativt sågs problem inom samtliga områden som NHP omfattar, d.v.s. fysisk rörlighet, emotionella reaktioner, energi, sömn, social isolering, smärta och ”NHP index of distress” (NHPD). Postoperativt sågs individuella förbättringar inom samtliga områden men alla patienter förbättrades inte inom alla områden. Ett eventuellt mönster bestående av förbättringar inom fysisk rörlighet, emotionella reaktioner, energi och NHPD kunde identifieras. Dessa resultat talar för att denna typ av utvärdering är värdefull vid transplantationsförsök vid Parkinsons sjukdom, att ny värdefull information erhålls som inte framkommmer i utvärdering enligt traditionella symptominriktade protokoll, och att transplantation av dopaminceller har potential att ge effekter utöver de motoriska konsekvenserna av sjukdomen.


Scratching feels better than thinking
8. ACKNOWLEDGEMENTS

Experience tells me that if you claim to have read the preceding pages of this book before winding up on these particular ones, either your character or your truthfulness can probably be questioned...

Someone recently drew a parallel between the PhD process and J.R.R. Tolkien’s The Lord of the Rings, with Frodo Baggins symbolizing the PhD-student and all the characters and adventures along his way the various people and events that are encountered by the PhD-student on his or her way to the doctoral degree. As much as I like this metaphor, there is another one that has come to mind during the final stages of my own process - childbirth... Just like Frodo’s adventures, there is no way I ever will be able to fully validate this metaphor through personal experience. However, as far as I can tell at this point, there appears to be several common aspects between the PhD process and bringing a new child into this world. Initially, you are not necessarily aware of what is going on. Then, after a while, you have your suspicions and after some formalities, including exposure of your past activities and current state, you receive a message that confirms what is going on. You are then pretty happy about the current state of affairs and begin making some vague plans regarding the process leading up to the event, as well as what is to come thereafter. Because anyway you slice it, if everything goes according to plans, you will never be quite the same after it is done (in the childbearing case, you will become a parent; in the case of being a PhD-student, you will have three additional letters after your name). Apart the fact that you may need to adjust or give up some of your normal habits, life pretty much goes on, although with some uncomfortable interruptions every now and then, and by time more and more people in your surrounding become aware what you are up to and acknowledge it. About halfway through the process, things become more apparent. Your social life starts changing, but thanks to your new situation you have made lots of new contacts with peers and the like that you never would have met or known otherwise. You also find yourself at various gatherings presenting your own and listening to others’ experiences, and discuss them openly in a fashion that probably very few who are not in, or have not been through, the process would really understand. By the end you realize that it is time... You know exactly what is to come and need to be done. But it hurts. Painful late nights and the taste of blood bring you closer and closer. Then, finally, it starts to move and you see that something actually comes out (in the case of childbirth, it is a child; in the case of producing a PhD-thesis, it is a bunch of papers coming out from the printer...). Then, before you really will be able to get a first idea of how it turned out, you submit your result for initial inspection of vital signs (in the case of childbirth, to the attending gynaecologist; in the case of producing a PhD-thesis, to your supervisor...). Once initial approval has been granted, all you can do really is to nurse and polish the thing to the best of your ability, and wait and hope for
the best... Slightly later on, there will be another inspection by several other experts in the field, where you also will be required to give an honest and accurate oral presentation of the whole process. There are a few noteworthy differences between childbirth and producing a PhD-thesis, though (I would like to think...). First, in the latter case you will hopefully not have to worry about breastfeeding, diapers, comforters etc. during the years to come. Second, regardless how good an idea it may seem like, you will not automatically get an extensive, paid, time off work to be able to be with and admire your thesis. Finally, as opposed to childbirth, there is not one single person, including yourself, that you can hold responsible for (or who can take credit of...) the result of the PhD process, which finally brings me to the real purpose of this section of the book.

There are so many people that have played some kind of part in this process that I am virtually bound to leave many out from these pages. All great colleagues, co-workers, and friends that I have met and interacted with ever since I first came to the Department of Neurology in Lund in 1989, at my brief period at the Department of Neurology in Malmö 1991-1992, and during the past years as a PhD-student, have contributed in one way or another; there are so many I want to thank and I hope I have not left too many out.

First and foremost, my deepest gratitude goes to my supervisor, Olle Lindvall. I could probably spend several pages on this “subject” - but I will not. Let me just say that the years I have had the fortune to work together with you have been unforgettable and enormously rewarding. Your scientific brilliance, strategic thinking, never-ending curiosity and openmindedness (in no particular order) are outstanding. Your way with the pen is well known but must still be acknowledged. My gratitude goes also to the support, trust, freedom, and responsibility you have provided me with during these years. Without that, I would never have been able to accomplish what I have, either in terms of this thesis or within other related areas of work. I also appreciate your sense of humor and the very good friendship that we have developed over the years, and I hope that we once again will find reason to do another round of peanut inspections and be able to enjoy a pint or two at the Bryanston Court Hotel, wondering whether the night to follow will bring a draft or a sauna, while discussing science, opera, Gary Moore, or just life...

The Section of Restorative Neurology, which has been my scientific (and during some periods of intense work almost permanent) home for the past several years, with a blend of people and activities that is truly enjoyable. Zaza Kokaia, Merab Kokaia, Avtandil “Avto” Nanobashvili, Andreas Arvidsson, Christine Ekdahl, Paul Mohapel, Fredrik Asztely, Istvan Ferencz, Barbara Ahlfors, Eskil Elmér, Johan Bengzon, Klara Olofsdotter, and our “hang-arounds” Anders Tingström, Johan Hellsten, and Malin Wilson at MPU. A special thank you goes to my room mates, Elin Larsson (you are
such a relief sometimes), Gunnar Andsberg (who also probably managed to influence my neurogenesis in a positive manner while writing up this thesis), and Monica Lundahl (one of the few among us with a “normal” job...), for bearing with me. My very best regards to you all - thank you all for a great atmosphere and lots of fun!

I also want to thank all the great and inspiring people whom I have had the honest pleasure of collaborating with throughout this work. The Lund group, with Anders Björklund, Stig Rehncrona, Patrik Brundin, Håkan Widner, Åsa Petersén, Oskar Hansson, Björn Gustavii, Cecilia Ehrstedt, and Pål Wølner-Hanssen; the London groups, with David Brooks, Paola Piccini, Paul Morrish, and Nicola Pavese at the Hammersmith Hospital, and Niall Quinn, Lesley Crabb, Annette Schrag, Gregor Wenning, Richard Brown, John Rothwell, and Marjan Jahanshahi at the National Hospital for Neurology and Neurosurgery at Queen Square; the Munich/Marburg group, with Wolfgang Oertel, Oliver Pogarell, Andreas Kupsch, and Per Odin (previously in Lund); and the Manchester group, with Stephen McKenna and Diane Whalley – it has been a true pleasure to work with you all! Håkan, your contribution, in particular to the dyskinesia rating scale project and as an expert movement disorder clinician whom I am fortunate to have been able to work with and learn from, has been invaluable. Steve, your calm insightfulness is admirable and I am always thankful for you showing a path that eventually led me to open new doors. David, your innovative curiosity is a never-ending source of admiration and your sense of humor is relieving. Niall, few possess the bland of warmth, proficiency, and perspicacity that you do. Wolfgang, your ability to cover virtually anything from neurotrophic factors to health economy is inspiring. Anders, your scientific skills, sharpness, and problem-solving capacities are admirable. Stig, what you cannot accomplish with a stereotaxic frame in the OR is probably not worthwhile in the first place. Per, in a way you were the one who talked me into this. Lesley, thanks for being a great friend and colleague (now, back to that book chapter...). Patrik, thanks for being the generous, funny, and inspiring person you are (time for another review soon...?).

A special gratitude goes to Kerstin Ulander and Ewa Wallin, without your visionary determination I am pretty confident I would not have been where I am today. In a different sense, the same can be said about Carol Moskowitz - you are probably responsible for more than you necessarily are aware of... Ingalill Rahm-Hallberg, the selection and timing of novel graduate courses that you have introduced at the Faculty of Medicine must be acknowledged - not only were they invaluable learning experiences, they also eventually led to my own current teaching activities at your department. Ben Wright and Mike Linacre at the MESA Psychometric Laboratory, University of Chicago - what can I say? Ben, the living legend who made Rasch’s work known and applicable, I will never forget how you took time to sit down to discuss measurement theory with me one sunny November afternoon in Chicago. Mike, thanks for listening and being patient with an eager but confused student during another couple of sunny days in Chicago.
I also wish to acknowledge the staff at the hospital Department of Neurology, whom I have been with one way or another for the past 13 years. Unfortunately, you are way to many for all to be mentioned. I do, however, want to mention those at the clinical movement disorder team: Christine Nilsson, Ann Dahl, Gudrun Sigurdardóttir, Birgitta Gunnarsson, Elisabeth Rasmusson, Gunilla Lundgren, Anne Wiberg, Gun Lindahl, Ingrid Dahlman, and Gisela Railo. In particular, I want to mention Martin Grabowski who, during recent years, has proven an inspiring scientific collaborator. A very special gratitude goes to Birgitta Sandlund and Jan Reimer - thanks for many thoughtful discussions, assistance with data collection, and for being great friends. And to Irena Westergren, thank you for your support.

Others that I have met and/or worked with and who have contributed one way or the other include: Angela Cenci, thank you for all interesting discussions and warm friendship; everyone else in “team Björklund”, in particular Malin Andersson, Martin Lundblad, Deniz Kirik, Carl Rosenblad, Christian Winkler, Ulrica Englund, and Anita Frank; Mia Emgård, Jenny Karlsson, Lena Larsson, Gabi Kaminski-Schierle, and others at the Section for Neuronal Survival; Geoff Norman, the physicist who winded up as statistician/health researcher at McMaster University and who is one of the few who can make a joke out of a standard deviation and get away with it - I hope we’ll meet soon again; Lisa Johnston, Julie Carter, Mickie Welsh, Orna Moore, Kirsten Turner, Peggy Gray, Elke Rost-Ruffner, Marlene Reimer, and Carol White, just to mention a few of the many inspiring international nurse research colleagues I have had the pleasure to meet; Herman Ubachs, Virginia Prendergast, Vicki Roach, and Barbara Lester of the Executive Committee, as well as all the Board members, whom I have had the fortune to work together with in the World Federation of Neuroscience Nurses; also, all those involved in the Movement Disorder Caring Society of Sweden, in particular the Board members Arja Höglund, Yvonne Jansson, Margareta Svensson, Viola Lundström, and Lena Quittberg, whom I still am trusted to work with; Meg Morris, the gait expert from Down Under - hope to see you soon again; Susanna Lindvall; Mary Baker; Ulf Persson; Sara Nordling; Per Nyberg; and the students that I have been trusted to supervise and from whom I learn a lot. A very special thanks to Bengt Mattsson for his unbeatable skills in scientific illustrations, generosity, and great sense of humor.

I also wish to thank the patients who have participated in this work for their cooperation. Without you these pages would have been fairly empty...

On a more personal, or perhaps I should say non-work related, level there are of course also several people to whom I am humbly thankful. First of all, of course, my parents, who always have believed in me and provided the best of support; my mother during my PhD years, and my mother and father before that. I love you! Ulrika, thank you for all your wonderful dinners and good friendship - it has often been a relief. To my fellow patres of PsLxV for providing a relaxed, and often very welcomed, forum of
life exploration. In a somewhat different sense, the same can be said to Pelle and Christel. To my dear, dear friends Tessie, Jim, VP, and JD (I am so sorry I missed this year’s Sunday brunch at the “Globe”...), and, of course, the official unofficial drinking and entertainment committee, Steve and Kate: I cannot wait to see you all again - whatever piece of the globe we may happen to set foot together next.
Don’t put it behind you – file it

III
9. REFERENCES


Backer JH. Stressors, social support, coping, and health dysfunctions in individuals with Parkinson’s disease. J Gerontol Nurs 2000; 7-16.


Marsden CD, Olanow CW. The causes of Parkinson’s disease are being unraveled and rational neuroprotective therapy is close to reality. *Ann Neurol* 1998; 44 (Suppl 1): S189-S196.


Stebbins GT, Goetz CG. Factor analysis of the motor section of the unified Parkinson’s disease rating scale during the off-state. *Mov Disord* 1999; 14: 585-589.


Stern G. Did parkinsonism occur before 1817? *J Neurol Neurosurg Psychiatry* 1989; Suppl: 11-12.


Strandberg M, Larsson L. Rating scales for parkinsonism and dyskinesias in squirrel monkeys. Faculty of Medicine, Lund University, 1995.


Whalley D. Recent advances in the evaluation of the scaling properties of health status and quality of life instruments. *Journal D’Economie Medicale* 1996; 14: 73.


Wright BD, Linacre JM. Observations are always ordinal; measurements, however, must be interval. *Arch Phys Med Rehabil* 1989; 70: 857-860.


