Outcome Measurement and Functional Prognosis in early Multiple Sclerosis

de Groot, V.

2007

Outcome Measurement and Functional Prognosis in early Multiple Sclerosis

Vincent de Groot

Outcome Measurement and Functional Prognosis in early Multiple Sclerosis Vincent de Groot

Outcome Measurement and Functional Prognosis in early Multiple Sclerosis

The studies presented in this thesis were carried out at the department of Rehabilitation Medicine of the VU University Medical Center in Amsterdam in close collaboration with the MS center of the VU University Medical Center and the EMGO institute of the VU University Medical Center.

The studies presented in this thesis were supported by a grant of the Netherlands Organization for Scientific Research (NWO 940-33-009).

The printing of this thesis was generously sponsored by Allergan, the Anna Fonds, Apotheekzorg, Basko, Biogen Idec, Coloplast, Kamer orthopedie, Livit orthopedie BV, Medtronic, Noppe orthopedie, Pfizer and Schering.

ISBN: 978-90-6464-155-8

Cover: Endeloos Grafisch Ontwerp (www.endeloos.nl)
Printing: Ponsen & Looijen BV, Wageningen
Lay-out: Jan Bruin (www.bred.nl)

VRIJE UNIVERSITEIT

Outcome Measurement and Functional Prognosis in early Multiple Sclerosis

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. L.M. Bouter,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Geneeskunde
op vrijdag 5 oktober 2007 om 10.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Vincent de Groot

geboren te Alkmaar

promotoren: prof.dr. L.M. Bouter
prof.dr. G.J. Lankhorst
prof.dr. C.H. Polman
copromotor: dr. H. Beckerman

Contents

1

Introduction

Vincent de Groot

CHAPTER 1

Introduction

Multiple Sclerosis (MS) is a chronic progressive autoimmune disease that is characterized by lesions, varying in time and place, in the Central Nervous System (CNS).[1] Because all parts of the CNS can be affected, a wide variety of symptoms can result from MS, including changes in sensation, visual problems, muscle weakness, depression, difficulties with co-ordination and speech, cognitive problems, and pain. In addition to these neurological symptoms, severe fatigue is often present. MS has an incidence of approximately 7 per 100,000 and a prevalence of approximately 120 per 100,000, and it affects twice as many women as men.[1] The disease usually starts between 20-40 years of age, which means that mainly young adults who have just started a career or a family will be affected.

Pathological changes, such as inflammation, demyelination and axonal loss, that are found in the CNS, are reflected in the different disease courses.[2] The exacerbations in Relapsing-Remitting MS (RRMS) can be attributed to inflammation. Demyelination and axonal loss are responsible for lasting neurological damage, and, therefore, are associated with accumulation of persistent neurological deficits. FIGURE 1.1 shows the different disease courses. FIGURE 1.1A shows that exacerbations in RRMS do not necessarily recover completely, FIGURE 1.1B shows that exacerbations can be present in Secondary Progressive MS (SPMS), and FIGURE 1.1C shows that progression in Primary Progressive MS (PPMS) is not

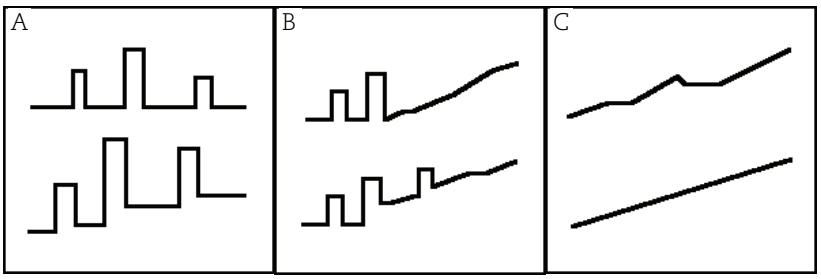


FIGURE 1.1 Different types of multiple sclerosis (A) Relapsing-Remitting. (B) Secondary Progressive. (C) Primary Progressive.

necessarily linear. At the start of the disease approximately 80% of the patients have an RRMS and 20% have an PPMS disease course. Eventually, in 60% of the patients with RRMS the disease course will change into SPMS.

The majority of studies focusing on the natural history of MS have used the Expanded Disability Status Scale (EDSS)[3,4] to measure disease severity.[5-11] However, this scale has been criticized because it has unsatisfactory validity, and its reliability is poor.[4,12,13] The EDSS combines the measurement of neurological deficits and mobility in one scale. Other relevant domains of functioning are not included in the EDSS. Therefore, little is known about the progression of disability in other domains of functioning in patients with MS.

In order to predict the progression of MS, several determinants of disease progression have been studied in the literature. Reviews of the studies that investigated determinants of the clinical course showed that a progressive disease course, higher age at the time of diagnosis, less than one year between relapses, and impairments of pyramidal or cerebellar tracts are all associated with an unfavourable disease course, whereas an exacerbation as first sign of MS, a high recovery rate after the first exacerbation, and afferent or monoregional symptoms are associated with a more favourable disease course.[5-11] A recent systematic review showed that the most robust predictors of long-term physical disability in RRMS are sphincter symptoms at onset, incomplete recovery from the first exacerbation, and a short interval between the first and second exacerbation.[14] However, this information on prognostic factors is derived from large group studies, and has not yet been shown to improve prognostication in individual patients.

Aims

The research that is described in this thesis was initiated in order to study these gaps in the currently available literature on MS. First of all, we were interested in the responsiveness (*i.e.* the ability of an

outcome measure to measure longitudinal changes) of several outcome measures that are frequently used in MS research. Secondly, we wanted to study the course of MS in other domains of functioning. Thirdly, we were interested in the determinants of changes in functioning. And finally, we wanted to investigate whether it is possible to accurately predict future functioning in individual patients.

Methods

We used the International Classification of Functioning (ICF) as theoretical framework in this thesis.[15] The ICF describes how patients live with their disease, and therefore is looking beyond mortality and disease. It is a classification of functioning that describes body functions, bodystructures, activities and participation, and takes personal and environmental factors into account. FIGURE 1.2 shows the theoretical relationships between the different factors. We also used the ICF to classify the outcome measures that were used in this thesis, to choose the domains of functioning that we were interested in, and to guide our analyses.

The outcomes of interest were divided into the following domains: disease severity, mobility, mental health, social function-

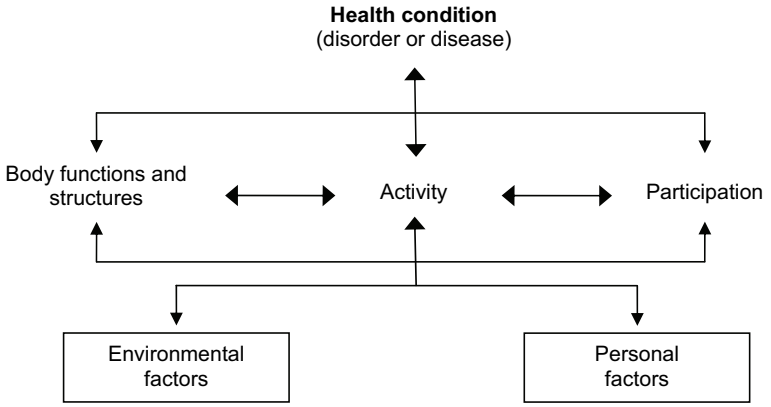


FIGURE 1.2 The International Classification of Functioning model and the theoretical relationships between its factors.

ing and general health. For the outcome assessments we used (sub-) scales of the EDSS,[3,16,17] the MS Functional Composite,[18-24] the Short and Graphic Assessment Scale,[25] the Action Research Arm Test (ARAT),[26,27] the Disability and Impact Profile (DIP),[28-32] the Functional Independence Measure (FIM),[33-35] the Rehabilitation Activities Profile (RAP),[36,37] the Rivermead Mobility Index (RMI),[38-41] and the Medical Outcome Study Short Form 36 (SF36).[42-45]

We measured potential determinants of functioning in the domains of personal and disease characteristics, basic functions, psychosocial characteristics, and basic activities. Although our main focus was on the disease MS, it is possible that the limitations in functioning are not only due to MS. Therefore, we also investigated whether relevant comorbid conditions determined functioning.

We included consecutive patients with a diagnosis of definite MS according to the Poser-criteria[46], and who visited the outpatient clinics of the participating neurology departments of the VU University Medical Center, the Erasmus Medical Center, the Academic Medical Center, the Sint Lucas-Andreas Hospital, and the OLVG Hospital. Patients with other neurological disorders, or systemic or malignant neoplastic diseases, were excluded. The measurements took place at baseline and at six months, and after one, two and three years. If a patient had a relapse, the measurements were postponed for a few weeks until the relapse had subsided. The patients were visited at home in order to minimize drop-out.

Outline of the thesis

Chapters 2 and 3 of this thesis describe the clinimetric studies. In CHAPTER 2 we critically review the clinimetric properties of the available comorbidity measures. In CHAPTER 3 we present a comprehensive analysis of the suitability of several outcome measures to evaluate longitudinal changes in MS. CHAPTER 4 describes the course of MS in the domains of neurological deficits, mobility, mental health, social functioning and general health. CHAPTER 5

contains a detailed study of the determinants of social dysfunctioning in the early stages of MS. In CHAPTER 6 we investigate whether it is possible to accurately predict future functioning. Finally, in CHAPTER 7 we will critically discuss the main issues related to the study design, the clinimetric studies, and the studies of the disease course.

References

1. Compston A, Coles A. Multiple sclerosis. Lancet 2002; 359:1221-1231.
2. Lublin FD, Reingold SC. Defining the clinical course of multiple sclerosis: results of an international survey. National Multiple Sclerosis Society (USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis. Neurology 1996; 46:907-911.
3. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology 1983; 33:1444-1452.
4. Hobart J, Freeman J, Thompson A. Kurtzke scales revisited: the application of psychometric methods to clinical intuition. Brain 2000; 123:1027-1040.
5. Runmarker B, Andersen O. Prognostic factors in a multiple sclerosis incidence cohort with twenty-five years of follow-up. Brain 1993; 116:117-134.
6. Weinshenker BG, Bass B, Rice GPA, Noseworthy J, Carriere W, Baskerville J *et al.* The natural history of multiple sclerosis: A geographically based study. 2. Predictive value of the early clinical course. Brain 1989; 112:1419-1428.
7. Weinshenker BG, Bass B, Rice GPA, Noseworthy J, Carriere W, Baskerville J *et al.* The natural history of multiple sclerosis: A geographically based study. I. Clinical course and disability. Brain 1989; 112:133-146.
8. Weinshenker BG, Rice GPA, Noseworthy JH, Carriere W, Baskerville J, Ebers GC. The natural history of multiple sclerosis: A geographically based study. 3. Multivariate analysis of predictive factors and models of outcome. Brain 1991; 114:1045-1056.
9. Amato MP, Ponziani G. A prospective study on the prognosis of multiple sclerosis. Neurol Sci 2000; 21:S831-S838.
10. Confavreux C, Vukusic S, Adeleine P. Early clinical predictors and progression of irreversible disability in multiple sclerosis: an amnesic process. Brain 2003; 126:770-782.
11. Pittock SJ, Mayr WT, McClelland RL, Jorgensen NW, Weigand SD, Noseworthy JH *et al.* Change in MS-related disability in a population-based cohort: a 10-year follow-up study. Neurology 2004; 62:51-59.
12. Noseworthy JH. Clinical scoring methods for multiple sclerosis. Ann Neurol 1994; 36:S80-S85.

13. Sharrack B, Hughes RA. The Guy's Neurological Disability Scale (GNDS): a new disability measure for multiple sclerosis. Mult Scler 1999; 5:223-233.
14. Langer-Gould A, Popat RA, Huang SM, Cobb K, Fontoura P, Gould MK *et al.* Clinical and demographic predictors of long-term disability in patients with relapsing-remitting multiple sclerosis: a systematic review. Arch Neurol 2006; 63:1686-1691.
15. WHO. International Classification of Functioning, Disability and Health: ICF. Geneva: WHO, 2001.
16. Whitaker JN, Mcfarland HF, Rudge P, Reingold SC. Outcomes assessment in multiple sclerosis clinical trials: a critical analysis. Mult Scler 1995; 1:37-47.
17. Rudick R, Antel J, Confavreux C, Cutter G, Ellison G, Fischer J *et al.* Clinical outcomes assessment in multiple sclerosis. Ann Neurol 1996; 40:469-479.
18. Cutter GR, Baier ML, Rudick RA, Cookfair DL, Fischer JS, Petkau J *et al.* Development of a multiple sclerosis functional composite as a clinical trial outcome measure. Brain 1999; 122:871-882.
19. Fischer JS, Rudick RA, Cutter GR, Reingold SC. The Multiple Sclerosis Functional Composite Measure (MSFC): an integrated approach to MS clinical outcome assessment. National MS Society Clinical Outcomes Assessment Task Force. Mult Scler 1999; 5:244-250.
20. Cohen JA, Fischer JS, Bolibrush DM, Jak AJ, Kniker JE, Mertz LA *et al.* Intrarater and interrater reliability of the MS functional composite outcome measure. Neurology 2000; 54:802-806.
21. Kalkers NF, de Groot V, Lazeron RH, Killestein J, Ader HJ, Barkhof F *et al.* MS functional composite: relation to disease phenotype and disability strata. Neurology 2000; 54:1233-1239.
22. Miller DM, Rudick RA, Cutter G, Baier M, Fischer JS. Clinical significance of the multiple sclerosis functional composite: relationship to patient-reported quality of life. Arch Neurol 2000; 57:1319-1324.
23. Kalkers NF, Bergers E, Castelijns JA, van Walderveen MA, Bot JC, Ader HJ *et al.* Optimizing the association between disability and biological markers in MS. Neurology 2001; 57:1253-1258.
24. Hoogervorst EL, van Winsen LM, Eikelenboom MJ, Kalkers NF, Uitdehaag BM, Polman CH. Comparisons of patient self-report, neurologic examination, and functional impairment in MS. Neurology 2001; 56:934-937.
25. Vaney C, Vaney S, Wade DT. SaGAS, the Short and Graphic Ability Score: an alternative scoring method for the motor components of the Multiple Sclerosis Functional Composite. Mult Scler 2004; 10:231-242.
26. Lyle RC. A performance test for assessment of upper limb function in physical rehabilitation treatment and research. Int J Rehabil Res 1981; 4:483-492.

27. van der Lee JH, Beckerman H, Lankhorst GJ, Bouter LM. The responsiveness of the Action Research Arm test and the Fugl-Meyer Assessment scale in chronic stroke patients. J Rehabil Med 2001; 33:110-113.
28. Laman H, Lankhorst GJ. Subjective weighting of disability: An approach to quality of life assessment in rehabilitation. Disabil Rehabil 1994; 16:198-204.
29. Jonsson A, Dock J, Ravnborg MH. Quality of life as a measure of rehabilitation outcome in patients with multiple sclerosis. Acta Neurologica Scandinavica 1996; 93:229-235.
30. Lankhorst GJ, Jelles F, Smits RCF, Polman CH, Kuik DJ, Pfennings LEMA *et al.* Quality of life in multiple sclerosis: The disability and impact profile (DIP). J Neurol 1996; 243:469-474.
31. Pfennings LE, van der Ploeg HM, Cohen L, Polman CH. A comparison of responsiveness indices in multiple sclerosis patients. Qual Life Res 1999; 8:481-489.
32. Cohen L, Pouwer F, Pfennings LE, Lankhorst GJ, van der Ploeg HM, Polman CH *et al.* Factor structure of the Disability and Impact Profile in patients with multiple sclerosis. Qual Life Res 1999; 8:141-150.
33. Granger CV, Cotter AC, Hamilton BB, Fiedler RC, Hens MM. Functional assessment scales: A study of persons with multiple sclerosis. Arch Phys Med Rehabil 1990; 71:870-875.
34. Kidd D, Howard RS, Losseff NA, Thompson AJ. The benefit of inpatient neurorehabilitation in multiple sclerosis. Clin Rehabil 1995; 9:198-203.
35. Marolf MV, Vaney C, Konig N, Schenk T, Prosiegel M. Evaluation of disability in multiple sclerosis patients: A comparative study of the Functional Independence Measure, the extended Barthel Index and the Expanded Disability Status Scale. Clin Rehabil 1996; 10:309-313.
36. van Bennekom CAM, Jelles F, Lankhorst GJ, Bouter LM. The Rehabilitation Activities Profile: A validation study of its use as a disability index with stroke patients. Arch Phys Med Rehabil 1995; 76:501-507.
37. van Bennekom CAM, Jelles F, Lankhorst GJ, Bouter LM. Responsiveness of the Rehabilitation Activities Profile and the Barthel Index. J Clin Epidemiol 1996; 49:39-44.
38. Collen FM, Wade DT, Robb GF, Bradshaw CM. The Rivermead Mobility Index: a further development of the Rivermead Motor Assessment. Int Disabil Stud 1991; 13:50-54.
39. Forlander DA, Bohannon RW. Rivermead Mobility Index: a brief review of research to date. Clin Rehabil 1999; 13:97-100.
40. Hsieh CL, Hsueh IP, Mao HF. Validity and responsiveness of the Rivermead Mobility Index in stroke patients. Scand J Rehabil Med 2000; 32:140-142.
41. Antonucci G, Aprile T, Paolucci S. Rasch analysis of the Rivermead Mobility

Index: a study using mobility measures of first-stroke inpatients. Arch Phys Med Rehabil 2002; 83:1442-1449.

42. Brunet DG, Hopman WM, Singer MA, Edgar CM, MacKenzie TA. Measurement of health-related quality of life in multiple sclerosis patients. Can J Neurol Sci 1996; 23:99-103.
43. Vickrey BG, Hays RD, Harooni R, Myers LW, Ellison GW. A health-related quality of life measure for multiple sclerosis. Qual Life Res 1995; 4:187-206.
44. Freeman JA, Hobart JC, Langdon DW, Thompson AJ. Clinical appropriateness: a key factor in outcome measure selection: the 36 item short form health survey in multiple sclerosis. J Neurol Neurosurg Psychiatry 2000; 68:150-156.
45. Hobart JC, Lamping DL, Freeman JA, Langdon DW, McLellan DL, Greenwood RJ *et al.* Evidence-based measurement: which disability scale for neurologic rehabilitation? Neurology 2001; 57:639-644.
46. Poser CM, Paty DW, Scheinberg L, McDonald WI, Davis FA, Ebers GC *et al.* New diagnostic criteria for multiple sclerosis: guidelines for research protocols. Ann Neurol 1983; 13:227-231.

2

How to measure comorbidity: a critical review of available methods

Vincent de Groot, Heleen Beckerman,
Gustaaf J. Lankhorst & Lex M. Bouter

J Clin Epidemiol 2003; 56:221-229

Abstract

Objectives. To systematically review available methods to measure comorbidity and to assess their validity and reliability.

Methods. A search was made in Medline and Embase, with the keywords comorbidity and multi-morbidity, to identify articles in which a method to measure comorbidity was described. The references of these articles were also checked, and using a standardized checklist the relevant data were extracted from these articles. An assessment was made of the content, concurrent, predictive and construct validity, and the reliability.

Results. Thirteen different methods to measure comorbidity were identified: one disease-count and 12 indexes. Data on content and predictive validity were available for all measures, while data on construct validity were available for nine methods, data on concurrent validity and inter-rater reliability for eight methods, and data on intra-rater reliability for three methods. The Charlson Index is the most extensively studied comorbidity index for predicting mortality. The Cumulative Illness Rating Scale (CIRS) addresses all relevant body systems without using specific diagnoses. The Index of Coexisting Disease (ICED) has a two-dimensional structure, measuring disease severity and disability, which can be useful when mortality and disability are the outcomes of interest. The Kaplan Index was specifically developed for use in diabetes research.

Conclusions. The Charlson Index, the CIRS, the ICED and the Kaplan Index are valid and reliable methods to measure comorbidity that can be used in clinical research. For the other indexes, insufficient data on the clinimetric properties are available.

Introduction

As early as 1970, Alvan Feinstein noted that 'the failure to classify and analyse comorbid diseases has led to many difficulties in medical statistics',[1] because comorbidity can affect the moment of detection, prognosis, therapy and outcome. Comorbidity can play an important role in different types of research. In etiological studies the relationship between comorbid conditions and an index disease can be investigated. Comorbidity can be the cause or the consequence of an index disease. It is also possible that the index disease and the comorbid conditions share the same risk factors. In diagnostic studies, comorbidity can obscure the relationship between the test under study and the index disease. In these fields of research it might be particularly useful to analyse every disease as a separate variable, in order to gain insight into the relationship between individual diseases and the index disease at issue. However, this method is not feasible in small studies, because of reduced efficiency of the analysis. Randomized controlled trials (RCTs) and prognostic studies can also be complicated by comorbidity. Comorbidity can either act as a confounder, threatening the internal validity, or as an effect modifier, threatening the internal and external validity of the study. For these purposes an efficient method is needed to measure comorbidity.

There are four important reasons for measuring comorbidity. The first reason is to be able to correct for confounding, and thus improve the internal validity of studies. The second reason is to be able to identify effect modification. A third reason is the desire to use comorbidity as a predictor of study outcome or natural history. Finally, a comprehensive comorbidity measure, including many co-occurring comorbid conditions in one valid variable, is needed for reasons of statistical efficiency.

Since an overview of available methods to measure comorbidity is still lacking, the following research question was formulated: Which methods are available for measuring comorbidity that can be used in RCTs and prognostic studies?

CHAPTER 2

Methods

A search was made in the electronic databases of Medline (from January 1966 to September 2000) and Embase (from January 1988 to September 2000). The following keywords were used to identify potentially useful articles: comorbidity, multi-morbidity, and co-existing disease. Articles in which the focus was on comorbidity assessment or comorbidity was an important (prognostic) variable were considered for inclusion. In the literature several terms are being used in comorbidity research. Also, there is no consensus regarding the definition of these terms. This review focuses on methods that can be used to assess the burden of diseases that exists besides an index disease, *i.e.* the disease under study. Subsequently, it was carefully checked whether the articles described methods to assess comorbidity or the clinimetric properties of these methods. Then, the reference lists of the retrieved articles were checked for other eligible articles. Methods that analyse every disease as a separate variable were excluded. Because we focus on comorbidity in medical patients, methods that deal with psychiatric comorbidity in populations in which the index disease (the main disease under study) is also psychiatric were also excluded.

An assessment was made of the content, criterion and construct validity, as well as the reliability of the identified methods.[2] As an indication of the administrative burden a description of the information that is needed to arrive at a score on the measure is given. Since the focus of this review is on comorbidity as a determinant and not on comorbidity as an evaluative measure responsiveness was not assessed. Content validity concerns the extent to which a measure includes all relevant items: it is a qualitative assessment. To describe content validity a short description is given of the items included in the method, whether or not some type of weighting or (pathophysiologic) severity ranking was applied, which information is needed to obtain a score, how to arrive at the final score, and whether or not adaptations of the method are available for specific purposes.

Streiner and Norman[2] define criterion validity as: 'the correlation of a scale with some other measure of the trait or disorder under study, ideally, a 'gold standard' (the criterion) which has been used and accepted in the field'. Unfortunately there is no 'gold standard' available for measuring comorbidity in medical patients, so one has to use other comorbidity measures for comparison. In this situation the decision on which measure is best depends not only on statistical tests but also on clinical judgement. Criterion validity can be subdivided into concurrent and predictive validity. Concurrent validity is assessed by correlating the measure under study with the criterion measure, which is given at the same time.[2] Parameters used to assess concurrent validity are the Spearman or Pearson correlation coefficients (r) and the Intra-class Correlation Coefficient (ICC). Although it is very difficult to determine cut-off points for correlation coefficients, because there are many factors influencing their value,[2] correlation coefficients exceeding 0.4 were considered to be moderate and those exceeding 0.75 were considered to be high.[3] Predictive validity[2] is the ability of a measure to predict future events or future scores on the outcome measure of interest. The assessment of predictive validity was based on parameters obtained from survival analysis, proportional hazards models and linear or logistic regression models. Points of interest were the relative risks (RR), relative hazards (RH), odds ratios (OR), explained variance (r^2) and the area under the receiver operating characteristic curve (AUC). If regression models predicting future events were significant, or significantly improved after adding the comorbidity measure under study, this was considered to support predictive validity.

The assessment of construct validity encompasses the testing of hypotheses regarding the relationship of the measure under study with other more or less related traits (constructs)[2], such as age, mortality, ADL, length of stay or number of medications taken. There are several methods that can be used to assess construct validity, such as correlation coefficients and comparing means or proportions in different populations. Whether or not construct validity is confirmed

will be discussed for every measure, because it is not possible to formulate comprehensive rules for assessing construct validity.

Different types of reliability were also investigated: test-retest reliability, and intra and inter-rater reliability. Parameters used to assess reliability are (in descending order of appropriateness)[2] Intra-class Correlation Coefficients (ICC), (weighted) Cohen's Kappa ((w)K), correlation coefficients (r) and percentage of reliability. Reliability coefficients are considered to be fair when they exceed 0.4 and moderate when they exceed 0.75.[3]

For every identified method the first author (VdG) screened all related articles for data regarding the clinimetric properties of that method, using a standardized data-extraction form. Every method was either classified as an 'index' or as a 'disease count'. Methods were classified as an 'index' if the authors used weights or (pathophysiologic) severity rankings for the conditions or dimensions included in the index. Methods were classified as a 'disease count' if the authors solely used an enumeration of the number of conditions present. Methods that did not already have a name were given the name of the first author and the category to which they were assigned. For example, if the first author's name is Schwarz and the method was classified as an index, the method is referred to as the Schwarz Index.

Results

Thirteen different methods to assess comorbidity were identified and presented in alphabetical order in TABLE 2.1: one disease-count and 12 indexes. Data on content and predictive validity were available for all measures, while data on construct validity were available for nine methods, data on concurrent validity for eight methods, data on inter-rater reproducibility for eight methods, and data on intra-rater reliability for three methods.

The *Burden of Disease (BOD)* index[4] consists of 59 weighted disease categories, selected on the basis of a literature review and a consensus meeting of three physicians and a nurse. To obtain the

score on the BOD Index, medical records over the previous year were reviewed, based on standardized guidelines assessing symptoms, complications and need for and complexity of therapy. Although the clinimetric data are obtained from one single article, they support the concurrent and predictive validity of the BOD.[4] Furthermore, the authors found a low, but positive relationship with the Katz Activities of Daily Living scale (Katz ADL), supporting construct validity.[4] There was no relationship with the Sickness Impact Profile (SIP), whereas a weak positive relationship would be expected. Inter-rater reliability was good (TABLE 2.1).[4]

The *Charlson* Index[5] is the most extensively studied comorbidity index[4-26]. The 19 diseases included in the index have been selected and weighted on the basis of the strength of their association with mortality. It has been adapted for use with ICD-9 databases,[7-13,15,22] for use with patients with amputations,[17] transformed into a questionnaire,[6] and combined with age to form an age-comorbidity index.[16] Four out of six comparisons with other indices of comorbidity yielded correlation coefficients exceeding 0.40, supporting concurrent validity.[4,18-22] Predictive validity was confirmed by finding many significant relationships of the Charlson index with various criterion outcomes, such as mortality, disability, readmissions and length of stay (TABLE 2.1).[5,7,10-14,16,17,19,22-25] All relationships with various kinds of variables showed some, although not perfect, correlations in the anticipated directions, supporting construct validity.[4,6,8,9,12,14,15,18-20,24,26] Test-retest reliability is good, and inter-rater reliability is moderate to good (except for one outlier with an ICC = 0.16).[6,14,18,20] It should be noted that emphasis has been laid on the ability of the index to predict mortality (TABLE 2.1).

The *Cumulative Illness Rating Scale (CIRS)*[27] rates 13, conceptually valid, body systems (supporting content validity) on a 5-point (pathophysiologic) severity scale. It has been slightly adapted to form the CIRS-G (CIRS geriatric),[28] for which guidelines to enhance reliability have been formulated. Criterion validity has been confirmed by showing high correlation coeffients when comparing

TABLE 2.1 Comorbidity measures: main characteristics and study populations.

Index	Items	Weights	Information needed	Final sc
BOD index	59 diseases	0 – Not present 1 – Inactive 2 – Mild 3 – Moderate 4 – Severe	Clinical assessment of symptoms, complications, need for and complexity of therapy	Sum of
Charlson index	19 conditions	0 – RR 1.2-1.5 1 – RR 1.5-2.5 2 – RR 2.5-3.5 3 – RR 3.5-4.5 6 – RR >6		Sum of
CIRS	13 body systems	0 – No impairment to 4 – Life-threatening impairment	Clinical judgment	Sum of
Cornoni-Huntley index		1 – No comorbidity 2 – Impaired vision or hearing 3 – Heart disease, stroke or diabetes 4 – Both levels 2 and 3	Not specified	1-4
Disease count	Single diseases	None	- Interview - Patient record - ICD-9 codes	Numbe disease score d the lim applied
DUSOI index	Every present health problem is rated on four domains: - Symptom level - Complication level - Prognosis without treatment - Treatability (=prognosis with treatment)	0-5	Clinical judgment	Using a scoring leading 0-100
Hallstrom index	Two domains: - CF consisting of 10 conditions	None	Interview	- CF nu present - SF nu sympto

Final score	Adaptations	Study populations (n)
Sum of weights	None	Long-stay nursing home patients (194)
Sum of weights	- ICD-9 - Age - Patients with amputations - Questionnaire	Breast cancer (685), HIV + (129), ICU (201), SCI (330), Stroke (106), Cancer (203), Several surgical procedures (>10,000), Heart disease (>10,000), Pneumonia (>10,000), Elective non-cardiac surgery with diabetes or hypertension (218), Lower limb amputees (24)
Sum of weights	None	Mixed inpatients (472), Mixed deceased patients (72), Elderly and geriatric outpatients (141+181), Mixed institutionalized long-term care patients (439), SCI (330), Cancer (203)
1-4	None	Hypertension and age 75 to 84 years (878)
Number of present diseases, maximal score depends on the limits that are applied	Several	SCI (330), HIV + (395), ICU (105), Breast cancer (>10,000), Myocardial infarction (>10,000), Asthma (>10,000), Appendicitis (>10,000), Abdominal hernia (>10,000), Diverticulitis (>10,000), Biliairy tract disease (>10,000), Low back pain (>10,000), Pneumonia (>10,000), Diabetes with complications (>10,000), Rehabilitation inpatients (>10,000)
Using a weighted scoring paradigm leading to a score of 0-100	None	Mixed primary care patients (414+1191)
- CF number of present conditions - SF number of present symptoms	None	Out of hospital ventricular fibrillation (282)

TABLE 2.1 Continued

Index	Items	Weights	Information needed	Final sc
	- SF consisting of 6 cardiac symptoms			- total 1
Hurwitz index	- No comorbidity - Non-disabling comorbidity - Disabling comorbidity		Not specified	None
ICED - DS	- 14 disease categories	- 1-5	- Symptoms, signs and laboratory tests	Using a paradig
- FS	- 10 functional areas	- 1-5	- Level of impairment	scores 1
Incalzi index	52 Conditions	Based on RR for mortality	- History - Physical examination - Routine laboratory data - ECG - Chest X-ray	Sum of
Kaplan index	Vascular or non-vascular disease	0 – Noncogent, easy to control or no comorbidity 1 – Slight decompensation of vital system or non-threatening chronic conditions 2 – Impaired vital system or potentially threatening chronic condition 3 – Recent full decompensation of vital system or life-threatening chronic condition	Clinical information	Accordi severe grades as 3
Liu index	38 conditions	0 (Not present) – 5 (Active rehabilitation contra-indicated)	Medical records	Sum of
Shwartz index	21 conditions	Regression coefficient from a model to predict costs	Medical records or data-bases with ICD-9 codes	Sum of present

BOD = Burden of Disease index; CF = Chronic Factor; CIRS = Cumulative Illness Rating Scale; DUSOI = Duke Severity of Illness; DS: Disease Severity; ECG = Electro Cardio Gram; FS = Functional Severity; HIV+ = Human Immunodeficiency Virus positive; ICD-9 = International Classification of Diseases version 9; ICED = Index of Coexistent Disease;

Final score	Adaptations	Study populations (n)
- total 1.67 x CF + SF		
None	None	Back-related problems (931)
Using a scoring paradigm leading to scores 1-4	None	Total hip replacement (356), Long-stay nursing home residents (194)
Sum of weights	Adding points for every decade over 75	Mixed geriatric and general medicine (370)
According to the most severe condition, two grades 2 are ranked as 3	Expanded with several diagnoses	Diabetes (188), Breast cancer (404)
Sum of weights	None	Stroke (106)
Sum of weights of present conditions	None	Mixed patients (4439): Stroke, Lung disease, Heart disease, Prostate disease, Hip and femur fracture, low back disorder

ICU = Intensive Care Unit; RR = Relative Risk; SCI = Spinal Cord Injury; SF = Symptom Factor. For more detailed information on the validity and reliability of the comorbidity measures please see the appendix.

CIRS scores based on autopsy (the gold standard) with those based on health histories and chart reviews.[29] The CIRS was correlated with 4 other measures of comorbidity. Three out of five correlation coefficients exceeded 0.40, supporting concurrent validity.[19,20,28] There is little evidence to support predictive validity.[19,30] Small to fair positive correlations in the anticipated directions have been found for other variables, such as medication usage, ADL, IADL and age, supporting construct validity.[20,28,31] Inter-rater and test-retest reliability are good (TABLE 2.1).[19,20,27,28]

The *Cornoni-Huntley* index was intended to be used in a study investigating hypertension and associated comorbid conditions.[32] Because the authors used data that were gathered in another study, only information on visual acuity, hearing ability, heart disease, stroke and diabetes was available. Based on this information, they constructed a 4-level comorbidity index. The limited data support the predictive validity with mortality as outcome (TABLE 2.1).[32]

Several authors studied comorbidity by simply *counting the number of diseases* that exist in addition to the index disease of a patient.[19,33-36] Although this method seems to be quite straightforward, substantial differences exist with regard to the definitions used to define a condition as comorbid. Some authors used ICD-9 codes to count the total number of comorbid conditions,[19] whereas others made up a list of carefully selected comorbid conditions and counted the number of these conditions that were present, using medical records or ICD-9-CM codes.[33-35] Gross *et al*.[36] defined a condition as being comorbid if it required treatment or had altered an organ function. Three out of five correlations with other comorbidity measures or severity of illness measures exceeded 0.40, supporting concurrent validity.[19,36] Evidence from analyses based on several different outcomes supports predictive validity.[33,35] Construct validity was studied by comparing scores in two different groups showing expected differences[19,33] and relationships with several other variables showing small but positive associations in the anticipated directions (TABLE 2.1).[34,36]

The *Duke Severity of Illness (DUSOI)* index[37] was developed to assess ambulatory primary care patients, based on patient records, but has also been modified so that it can be used in direct contact between patient and clinician. First, all health problems are identified. For every health problem four domains (symptoms, complications, prognosis without treatment and treatability) are rated on a 5-point scale. The data support concurrent,[37] predictive[38] and construct[37] validity. Test-retest[37] and inter-rater reliability[37-39] are fair, and intra-rater reliability is fair to good[37-39] (TABLE 2.1).

The *Hallstrom* index was specifically developed to assist in predicting the outcome of cardiac arrest.[40] It consists of a chronic factor (CF) and a symptom factor (SF). The CF is the number of present conditions from a set of 10 conditions, and the SF is the number of present symptoms from a set of 6 symptoms related to cardiac disease. A low rank correlation between the two factors was found (0.22, $p < 0.001$), suggesting that the two scales assess two different concepts. The limited available data provide some support for predictive and construct validity (TABLE 2.1).[40]

The *Hurwitz* index was used in a study that was designed to assess the influence of comorbidity on the type of care (primary, specialist, chiropractic or other) that patients seek for their back problems[41]. Every patient was classified as having either no comorbidity, non-disabling comorbidity or disabling comorbidity. The index was only able to distinguish between medical and chiropractic care, thus providing limited support for predictive validity (TABLE 2.1).[41]

The *Index of Co-existent Disease (ICED)*[42] consists of two different dimensions, one measuring the disease severity of 14 categories of comorbid diseases (ICED-DS), and one measuring the 'overall functional severity' (disability) caused by comorbidity (ICED-FS). Scores are based on an explicit list of symptoms, signs and laboratory tests. All information contained in a medical chart can be used to calculate a score. The data support concurrent[4,21,42] and predictive validity.[42] Intra-rater reliability is good, and inter-rater reliability is fair (TABLE 2.1).[42,43]

The *Incalzi* index consists of 52 conditions, each weighted according to its relative risk (RR) for mortality.[44] An Incalzi age-index can be computed by adding 2, 3 or 4 points to the score of patients aged 76 to 85 years, 86 to 95 years and over 95 years, respectively.[44] Predictive validity was shown for both the dichotomised (cut-off values were identified on ROC curves) Incalzi and the Incalzi age-index for predicting mortality. Showing that the mortality rate for patients with scores above the 75th percentile was higher than for those with scores under the 75th percentile provide further support.[44] According to the authors, inter-rater reliability was good (data not presented in their article. TABLE 2.1).[44]

The *Kaplan* index uses two forms of classification, focussing on the type of comorbidity and the pathophysiologic severity of the present comorbid conditions, respectively. The type of comorbidity can be classified as vascular (hypertension, cardiac disorders, peripheral vascular disease, retinopathy and cerebrovascular disease) or non-vascular (lung, liver, bone and non-diabetic renal diseases). Pathophysiologic severity is rated on a 4-point scale, ranging from 0 (no, or easy to control comorbidity) to 3 (recent full decompensation of comorbid condition). The rating of the most severe condition determines the overall comorbidity score. Scores for vascular and non-vascular comorbidity can be calculated, based on the most severe condition in each sub-scale. There are two adaptations, the Modified Medical Comorbidity Index (MMCI) and the Adult Comorbidity Evaluation 27 (ACE-27), available.[45,46] Although the adaptations look promising, there are, to our knowledge, no articles published in which a detailed description of either the (content) validity or the reliability is given. The ability of the Kaplan index to predict mortality was studied.[5,14,47] The results support predictive validity (TABLE 2.1).

The *Liu* index consists of 38 conditions, and was specifically constructed for use in stroke outcome research.[18] Every condition is rated on a 6-point scale, ranging from 0 (not present) to 5 (active rehabilitation is contra-indicated). The Liu index has been com-

pared with the Charlson index, yielding a borderline fair correlation.[18] This result provides some support for concurrent validity, because, given the different objective of the Charlson index, this correlation should neither be too low nor too strong. The Liu index is able to predict scores on the Functional Independence Measure (FIM) and Length of Stay (LOS), supporting predictive validity.[18] Fair correlation coefficients in the anticipated directions between the Liu index and some other variables provide support for construct validity.[18,48] Inter-rater reliability is good (TABLE 2.1).[18]

The *Shwarz* index consists of 21 weighted conditions, selected from 52 conditions that were derived from the literature on the basis of their positive relationship with mortality or their negative influence on the treatment of the primary condition, using a regression model that was made to predict costs.[49] The Shwarz index can be used with medical records and with databases that use ICD-9-CM codes. Data supporting predictive validity were obtained from regression models predicting costs for the Shwartz index,[49] using sub-group analyses and other data-sets.[49] Models based on data from medical records performed better than models based on ICD-9-CM codes.[49] According to the authors, intra-rater reliability was high (data not presented in their article. TABLE 2.1).[49]

Discussion

Measuring comorbidity is an aspect of research that is receiving increasing attention in the literature. Several authors have discussed and compared the use of various selected methods to measure comorbidity.[50-53] This review describes methods that can be used to measure comorbidity in clinical research, without limiting the focus to certain index diseases or diagnostic groups. Thirteen different methods were identified. Six indexes used a carefully developed list of clearly defined diagnoses (BOD, Charlson Index, Hallstrom Index, Incalzi Index, Liu Index and Shwartz Index). Three indexes rated comorbidity burden by using a system that assessed the effect of comorbid conditions on spe-

cific body systems (CIRS, ICED and Kaplan Index). Two indexes rated comorbidity on a three or four point scale using very broad categories (Cornoni-Huntley Index and Hurwitz Index). Two methods used every present condition to calculate a score: one simply counted the number of present comorbid conditions (Disease count) and the other calculated a summary score based on weighted scores for every present comorbid condition (DUSOI).

Although all these methods were developed to measure comorbidity, in the current literature there is no consensus regarding the definition of comorbidity. According to Feinstein,[1] comorbidity is defined as 'any distinct additional entity that has existed or may occur during the clinical course of a patient who has the index disease under study'. Another definition of comorbidity is the co-occurrence of multiple diseases in one person.[51] In this respect, Van den Akker *et al.*[54] made a useful distinction between, on the one hand, multimorbidity (*i.e.* the co-occurrence of multiple chronic or acute diseases and medical conditions in one person) and, on the other hand, comorbidity as defined by Feinstein.[1] By definition, for research on multi-morbidity no index disease is used, whereas for comorbidity research an index disease is obligatory. Some of the methods identified can be used for both purposes, depending on whether the focus is on measuring the total burden of diseases in a patient (generic multi-morbidity measures) or the burden of comorbid diseases in addition to the condition of interest (generic comorbidity measures). In the latter case the index disease is omitted from the comorbidity measurement. Other methods, such as the Kaplan, the Liu, the Cornoni-Huntley and the Hallstrom indexes, were specifically developed to measure comorbid diseases in addition to one specific index disease (disease specific comorbidity measures). The other nine methods are generic measures.

Comorbidity indexes first identify the present comorbid diseases and subsequently apply weights or (pathophysiologic) severity ratings for these diseases. The technique of applying weights or

(pathophysiologic) severity ratings is very valuable. There is evidence that correcting for comorbidity by simply counting the number of existing diseases leads to another conclusion, than correcting for comorbidity by comorbidity indexes that use weights or (pathophysiologic) severity ratings.[55]

While there is a growing body of evidence that comorbidity, as a disease-count or an index, is an independent predictor of several outcomes,[33,35,56] relatively little is known about the effect of individual disease combinations on the outcome of interest.[50,52,53] A few authors studied the effect of combinations of individual diseases on disability.[56,57] They showed that the effects of some disease combinations on disability were additive, whereas the effects of other disease combinations were synergistic, leading to more disability than would be expected on the basis of addition. For this type of research, which studies the prognosis of comorbidity and multi-morbidity, large numbers of patients are required. Continuing this research to include several index diseases and outcomes could lead to the identification of comorbid conditions that are particularly relevant for one specific index disease and the chosen outcome. These comorbid conditions can subsequently be used to develop disease-specific comorbidity measures.

The development of a comorbidity measure is influenced by the population and outcome used.[52] For the Charlson index it was shown that other weights would have been applied if it had been developed for a different population.[8,9,12] Weights were based on the relative risk of dying, and were used to indicate that not all comorbid conditions have the same impact on the total comorbidity burden. What would have happened if they had chosen another outcome measure? It is likely that the weights would have been very different. Take, for example, osteoarthritis. Weights derived from regression analysis, using mortality as outcome will probably be very different from those using mobility as outcome. These influences should be taken into account when selecting an appropriate comorbidity measure.

Commonly used methods to obtain data that can be used to score comorbidity are interviews, questionnaires, physical examinations, medical chart reviews and coded databases. The completeness of data obtained from interviews and questionnaires depends on the ability of patients to adequately recall the diseases they suffer from. This ability is strongly influenced by the knowledge and memory of the patient.[6] Although the source of information, *i.e.* the patient, is the same for interviews and questionnaires, the correlation between scores based on interviews and questionnaires ranges from 0.45 to 0.63.[6,52] One advantage of these two methods is that they are easy to apply in settings in which there is no access to detailed patient records. Medical chart reviews probably yield the most complete data,[22,58-60] provided that all charts that exist for one patient are collected. Collecting all the charts and screening the content for relevant data can be rather time-consuming, thus increasing the administrative burden. The usefulness of coded databases for the assessment of comorbidity has been the subject of several articles.[59,61-63] A major problem is the limited space available for recording present diseases, and when there are multiple diagnoses, a selection must be made. More serious diseases, the disease for which the patient was admitted, and complications during hospitalisation have a higher chance of being recorded than chronic conditions,[15,59] introducing substantial bias. Increasing the number of diagnoses might limit this bias, although it is doubtful whether this will solve the problem.[59] Studies comparing scores derived from medical records with those derived from large ICD-9 code-based administrative databases, showed that data derived from medical records are more complete than those derived from administrative databases, especially with regard to asymptomatic diseases.[14,22,52,59,62] Administrative databases yield data for large patient groups, but for smaller studies data from medical records should preferably be used, although data from interviews or questionnaires are a useful alternative.

In conclusion, the Charlson Index, the CIRS, the ICED and the

Kaplan Index are valid and reliable methods to measure comorbidity that can be used in clinical research. For the other indexes, insufficient data on the clinimetric properties are available to assess their validity and reliability. When mortality is the outcome of interest, the Charlson Index has been studied most extensively, and there are several adaptations available. An advantage of the CIRS is the close resemblance to common clinical practice: it is structured according to clinically relevant body systems and uses a clear severity ranking that is clinically sound. Given the good validity and reliability, the CIRS seems a very useful comorbidity measure in clinical research. The ICED is the only measure included in this review that has a two-dimensional structure, measuring both pathophysiologic disease severity and disability. This might be particularly useful in studies in which mortality and disability are the outcomes of interest. The Kaplan index was specifically developed for use in diabetes research and contains clinically relevant information. It makes a distinction between vascular and non-vascular comorbidity and uses severity rankings based on parameters derived from common clinical practice. This good face validity, together with the good psychometric properties, makes the Kaplan Index a useful comorbidity index in clinical diabetes research.

References

1. Feinstein AR. The pre-therapeutic classification of co-morbidity in chronic disease. Journal of Chronic Diseases 1970; 23:455-468.
2. Streiner DL, Norman GR. Health Measurement Scales. A practical Guide to their Development and Use. 2, 1 -230. 1995. Oxford, Oxford University Press.
3. Fleiss JL. Reliability of measurement. The Design and Analysis of Clinical Experiments. New York: John Wiley & Sons, 1986: 1-32.
4. Mulrow CD, Gerety MB, Cornell JE, Lawrence VA, Kanten, DN. The relationship between disease and function and perceived health in very frail elders. J Am Geriatr Soc 1994; 42:374-380.
5. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. Journal of Chronic Diseases 1987; 40:373-383.
6. Katz JN, Chang LC, Sangha O, Fossel AH, Bates DW. Can comorbidity be meas-

ured by questionnaire rather than medical record review? Med Care 1996; 34:73-84.

7. Roos LLJ, Nicol JP, Cageorge SM. Using administrative data for longitudinal research: comparisons with primary data collection. Journal of Chronic Diseases 1987; 40:41-49.
8. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. J Clin Epidemiol 1992; 45:613-619.
9. Romano PS, Roos LL, Jollis JG. Adapting a clinical comorbidity index for use with ICD-9-CM administrative data: differing perspectives. J Clin Epidemiol 1993; 46:1075-1079; discussion 1081-1089.
10. D'Hoore W, Sicotte C, Tilquin C. Risk adjustment in outcome assessment: the Charlson comorbidity index. Methods Inf Med 1993; 32:382-387.
11. D'Hoore W, Bouckaert A, Tilquin C. Practical considerations on the use of the Charlson comorbidity index with administrative data bases. J Clin Epidemiol 1996; 49:1429-1433.
12. Ghali WA, Hall RE, Rosen AK, Ash AS, Moskowitz MA. Searching for an improved clinical comorbidity index for use with ICD-9-CM administrative data. J Clin Epidemiol 1996; 49:273-278.
13. Cleves MA, Sanchez N, Draheim M. Evaluation of two competing methods for calculating Charlson's comorbidity index when analyzing short-term mortality using administrative data. J Clin Epidemiol 1997; 50:903-908.
14. Newschaffer CJ, Bush TL, Penberthy LE, Bellantoni M, Helzlsour K, Diener-West M. Does comorbid disease interact with cancer? An epidemiologic analysis of mortality in a cohort of elderly breast cancer patients. J Gerontol 1998; Series A, Biological Sciences & Medical Sciences. 53:M372-M378.
15. Librero J, Peiro S, Ordinana R. Chronic comorbidity and outcomes of hospital care: length of stay, mortality, and readmission at 30 and 365 days. J Clin Epidemiol 1999; 52:171-179.
16. Charlson M, Szatrowski TP, Peterson J, Gold J. Validation of a combined comorbidity index. J Clin Epidemiol 1994; 47:1245-1251.
17. Melchiorre PJ, Findley T, Boda W. Functional outcome and comorbidity indexes in the rehabilitation of the traumatic versus the vascular unilateral lower limb amputee. Am J Phys Med Rehabil 1996; 75:9-14.
18. Liu M, Domen K, Chino N. Comorbidity measures for stroke outcome research: a preliminary study. Arch Phys Med Rehabil 1997; 78:166-172.
19. Rochon PA, Katz JN, Morrow LA, McGlinchey-Berroth R, Ahlquist MM, Sarkarati M *et al.* Comorbid illness is associated with survival and length of hospital stay in patients with chronic disability. A prospective comparison of three comorbidity indices. Med Care 1996; 34:1093-1101.

20. Extermann M, Overcash J, Lyman GH, Parr J, Balducci L. Comorbidity and functional status are independent in older cancer patients. J Clin Oncol 1998; 16:1582-1587.
21. Gabriel SE, Crowson CS, O'Fallon WM. A comparison of two comorbidity instruments in arthritis. J Clin Epidemiol 1999; 52:1137-1142.
22. Kieszak SM, Flanders WD, Kosinski AS, Shipp CC, Karp H. A comparison of the Charlson comorbidity index derived from medical record data and administrative billing data. J Clin Epidemiol 1999; 52:137-142.
23. West DW, Satariano WA, Ragland DR, Hiatt RA. Comorbidity and breast cancer survival: a comparison between black and white women. Ann Epidemiol 1996; 6:413-419.
24. Skiest DJ, Rubinstien E, Carley N, Gioiella L, Lyons R. The importance of comorbidity in HIV-infected patients over 55: a retrospective case-control study. Am J Med 1996; 101:605-611.
25. Poses RM, McClish DK, Smith WR, Bekes C, Scott WE. Prediction of survival of critically ill patients by admission comorbidity. J Clin Epidemiol 1996; 49:743-747.
26. Beddhu S, Bruns FJ, Saul M, Seddon P, Zeidel ML. A simple comorbidity scale predicts clinical outcomes and costs in dialysis patients. Am J Med 2000; 108:609-613.
27. Linn BS, Linn MW, Gurel L. Cumulative illness rating scale. J Am Geriatr Soc 1968; 16:622-626.
28. Miller MD, Paradis CF, Houck PR, Mazumdar S, Stack JA, Rifai AH *et al.* Rating chronic medical illness burden in geropsychiatric practice and research: application of the Cumulative Illness Rating Scale. Psychiatry Res 1992; 41:237-248.
29. Conwell Y, Forbes NT, Cox C, Caine ED. Validation of a measure of physical illness burden at autopsy: the Cumulative Illness Rating Scale. J Am Geriatr Soc 1993; 41:38-41.
30. Waldman E, Potter JF. A prospective evaluation of the cumulative illness rating scale. Aging 1992; 4:171-178.
31. Parmelee PA, Thuras PD, Katz IR, Lawton MP. Validation of the Cumulative Illness Rating Scale in a geriatric residential population. J Am Geriatr Soc 1995; 43:130-137.
32. Cornoni-Huntley JC, Foley DJ, Guralnik JM. Co-morbidity analysis: a strategy for understanding mortality, disability and use of health care facilities of older people. Int J Epidemiol 1991; 20:S8-17.
33. Page-Shafer K, Delorenze GN, Satariano WA, Winkelstein W Jr. Comorbidity and survival in HIV-infected men in the San Francisco Men's Health Survey. Ann Epidemiol 1996; 6:420-430.

34. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Med Care 1998; 36:8-27.
35. Stineman MG, Escarce JJ, Tassoni CJ, Goin JE, Granger CV, Williams SV. Diagnostic coding and medical rehabilitation length of stay: their relationship. Arch Phys Med Rehabil 1998; 79:241-248.
36. Gross PA, Stein MR, van Antwerpen C, DeMauro PJ, Boscamp JR, Hess W *et al.* Comparison of severity of illness indicators in an intensive care unit. Arch Intern Med 1991; 151:2201-2205.
37. Parkerson GR Jr, Broadhead WE, Tse CK. The Duke Severity of Illness Checklist (DUSOI) for measurement of severity and comorbidity. J Clin Epidemiol 1993; 46:379-393.
38. Parkerson GR Jr, Broadhead WE, Tse CK. Health status and severity of illness as predictors of outcomes in primary care. Med Care 1995; 33:53-66.
39. Parkerson GR Jr, Bridges-Webb C, Gervas J, Hofmans O, Lamberts H, Froom J *et al.* Classification of severity of health problems in family/general practice: an international field trial. Fam Pract 1996; 13:303-309.
40. Hallstrom AP, Cobb LA, Yu BH. Influence of comorbidity on the outcome of patients treated for out-of-hospital ventricular fibrillation. Circulation 1996; 93:2019-2022.
41. Hurwitz EL, Morgenstern H. The effects of comorbidity and other factors on medical versus chiropractic care for back problems. Spine 1997; 22:2254-63.
42. Greenfield S, Apolone G, McNeil BJ, Cleary PD. The importance of co-existent disease in the occurrence of postoperative complications and one-year recovery in patients undergoing total hip replacement. Comorbidity and outcomes after hip replacement. Med Care 1993; 31:141-154.
43. Imamura K, Mckinnon M, Middleton R, Black N. Reliability of a comorbidity measure: the Index of Co-Existent Disease (ICED). J Clin Epidemiol 1997; 50:1011-1016.
44. Incalzi RA, Capparella O, Gemma A, Landi F, Bruno E, Di Meo F *et al.* The interaction between age and comorbidity contributes to predicting the mortality of geriatric patients in the acute-care hospital. J Intern Med 1997; 242:291-298.
45. Piccirillo JF. Impact of comorbidity and symptoms on the prognosis of patients with oral carcinoma. Arch Otolaryngol Head Neck Surg 2000; 126:1086-1088.
46. Piccirillo JF. Importance of comorbidity in head and neck cancer. Laryngoscope 2000; 110:593-602.
47. Kaplan MH, Feinstein AR. The importance of classifying initial co-morbidity in evaluatin the outcome of diabetes mellitus. Journal of Chronic Diseases 1974; 27:387-404.
48. Liu M, Tsuji T, Tsujiuchi K, Chino N. Comorbidities in stroke patients as assessed with a newly developed comorbidity scale. Am J Phys Med Rehabil

1999; 78:416-424.
49. Shwartz M, Iezzoni LI, Moskowitz MA, Ash AS, Sawitz E. The importance of comorbidities in explaining differences in patient costs. Med Care 1996; 34:767-782.
50. Gijsen R, Hoeymans N, Schellevis FG, Ruwaard D, Satariano WA, van den Bos GA. Causes and consequences of comorbidity: a review. J Clin Epidemiol 2001; 54:661-674.
51. Guralnik JM. Assessing the impact of comorbidity in the older population. Ann Epidemiol 1996; 6:376-380.
52. Extermann M. Measurement and impact of comorbidity in older cancer patients. Crit Rev Oncol Hematol 2000; 35:181-200.
53. Extermann M. Measuring comorbidity in older cancer patients. Eur J Cancer 2000; 36:453-471.
54. van den Akker M, Buntinx F, Metsemakers JF, Roos S, Knottnerus JA. Multimorbidity in general practice: prevalence, incidence, and determinants of co-occurring chronic and recurrent diseases. J Clin Epidemiol 1998; 51:367-375.
55. Concato J, Horwitz RI, Feinstein AR, Elmore JG, Schiff SF. Problems of comorbidity in mortality after prostatectomy. JAMA 1992; 267:1077-1082.
56. Verbrugge LM, Lepkowski JM, Imanaka Y. Comorbidity and its impact on disability. Milbank Q 1989; 67:450-484.
57. Fried LP, Bandeen-Roche K, Kasper JD, Guralnik JM. Association of comorbidity with disability in older women: the Women's Health and Aging Study. J Clin Epidemiol 1999; 52:27-37.
58. Malenka DJ, McLerran D, Roos N, Fisher ES, Wennberg JE. Using administrative data to describe casemix: a comparison with the medical record. J Clin Epidemiol 1994; 47:1027-1032.
59. Iezzoni LI, Foley SM, Daley J, Hughes J, Fisher ES, Heeren T. Comorbidities, complications, and coding bias. Does the number of diagnosis codes matter in predicting in-hospital mortality? JAMA 1992; 267:2197-2203.
60. Iezzoni LI, Ash AS, Coffman GA, Moskowitz MA. Predicting in-hospital mortality. A comparison of severity measurement approaches. Med Care 1992; 30:347-359.
61. Wang PS, Walker A, Tsuang M, Orav EJ, Levin R, Avorn J. Strategies for improving comorbidity measures based on Medicare and Medicaid claims data. J Clin Epidemiol 2000; 53:571-578.
62. Humphries KH, Rankin JM, Carere RG, Buller CE, Kiely FM, Spinelli JJ. Co-morbidity data in outcomes research: are clinical data derived from administrative databases a reliable alternative to chart review? J Clin Epidemiol 2000; 53:343-349.
63. Iezzoni LI. The risks of risk adjustment. JAMA 1997; 278:1600-1607.

APPENDIX Data on validity and reliability of comorbidity measures. Short description of the measures (content validity), data on concurrent validity, selected significant results

Content validity	Concurrent validity	Predictive validity	Construc
BOD index			
Items: - 59 diseases *Weights:* 0 – not present 1 – inactive 2 – mild 3 – moderate 4 – severe *Information needed:* - clinical assessment of symptoms, complications, need for and complexity of therapy *Final score:* - sum of weights	Correlation with: - Charlson index = 0.43 - ICED = 0.55	Multivariate regression models predicting: - ADL: r^2 increases from 0.22 to 0.37 - SIP: r^2 increases from 0.39 to 0.48	Correlat - SIP: ab - ADL =
Refs[4]	Refs[4]	Refs [4]	Refs [4]
Charlson index			
Items: - 19 conditions *Weights:* 0 – RR 1.2-1.5 1 – RR 1.5-2.5 2 – RR 2.5-3.5 3 – RR 3.5-4.5 6 – RR >6 *Final score:* - sum of weights *Adaptations:* ICD-9 - age - patients with amputations - questionnaire	Correlation with: - BOD = 0.34 - ICED = 0.58, 0.69 - Liu index = 0.40 - CIRS = 0.51, 0.39 - ICD- 9 count = 0.25 Comparing adaptations: - K = 0.35-1.0 - r = 0.47	Multivariate models using the following dependent variables: - mortality (ST and LT) - disability - readmissions - LOS - RR = 1.38-2.31 - r^2 = 0.02-0.69 - $r_{observed/expected\ outcome}$ = 0.12-0.74 - AUC = 0.60-0.87	Expected - age - compli - mortal - blood t - nursing - LOS - charge - ADL - numbe - prescri - laborat - consul
Refs[5- 17]	Refs [4,6,9,12,14,18-22]	Refs [5,7,10-14,16,17,19,22-25]	Refs [6,8,

regarding predictive validity, tested relationships with other variables (construct validity), data on reliability and study populations.

Construct validity	Reliability	Study populations (n)
Correlation with: - SIP: absent - ADL = 0.21	Inter-rater: - ICC = 0.85	- Long-stay nursing home patients (194)
Refs [4]	Refs [4]	
Expected relationships with: - age - complications - mortality - blood transfusions - nursing home discharge - LOS - charges - ADL - number of hospitalizations - prescriptions - laboratory studies - consultations	Test-retest: - ICC = 0.91, 0.92 - R = 0.86, 0. 73, 0.94 Inter-rater: - ICC = 0.16 - WK = 0.95, 0.96 - r = 0.74	- Breast cancer (685) - HIV + (129) - ICU (201) - SCI (330) - Stroke (106) - Cancer (203) - Several surgical procedures (>10,000) - Heart disease (>10,000) - Pneumonia (>10,000) - Elective non-cardiac surgery with diabetes or hypertension (218) - Lower limb amputees (24)
Refs [6,8,15,18,20,24,26]	Refs [6,14,18,20]	

APPENDIX - Continued

Content validity	Concurrent validity	Predictive validity	Construc
CIRS			
Items:	Correlation:	Multivariate regression models	Compari
- 13 body systems	- autopsy data with	predicting:	with ill p
Weights:	patient record data = 0.75	- LOS: r^2 = 6.2%	had sign
(141+181)			
0 (no impairment) to	Correlation with:	- mortality: significant	
4 (life-threatening	- Carlson Index = 0.51, 0.39		Correlat
impairment)	- ICD-9 counts = 0.58		- medica
Information needed:	- global comorbidity rating		- ADL =
- clinical judgment	by a specialist = 0.48		- IADL =
Final score:	- global comorbidity rating		- age = 0
- sum of weights	by depressed patients = -0.30		
Adaptations:			
- none			
Refs [27]	Refs [19,20,28,29]	Refs [19,30]	Refs [20,2
Cornoni-Huntley index			
Item/weight:		Mortality rates for increasing	
1 – no comorbidity		scores in two populations:	
2 – impaired vision or hearing		score Iowa Boston	
3 – heart disease, stroke or diabetes		1 6% 8%	
4 – both levels 2 and 3		2 14% 12%	
		3 13% 18%	
		4 18% 25%	
Information needed:			
- not specified			
Final score:			
- 1-4			
Adaptations:			
- None			
Refs [32]		Refs [32]	
Disease count			
Items:	Correlations with:	Multivariate regression models	- Compa
- single diseases	- CIRS = 0.58	predicting:	with and
Weights:	- Charlson index = 0.25	- LOS: r^2 increases from 0.32 to 0.33	- Survivo
- none	- Computerised severity	- time to diagnosis: RH = 1.2, ns	5.47 vs.
	score = 0.42	- survival: ns	

Construct validity	Reliability	Study populations (n)
Comparing scores of healthy with ill populations: healthy had significantly lower scores	Test-retest: - r = 0.95	- Mixed inpatients (472) - Mixed deceased patients (72) - Elderly and geriatric outpatients
Correlation with: - medication usage = 0.31 - ADL = -0.47, 0.18, 0.58 - IADL = 0.23, 0.34 - age = 0.15, 0.45	Inter-rater: - ICC = 0.78-0.88 - r = 0.76, 0.80 - $W_{Kendall}$ = 0.83-0.91	- Mixed institutionalized long-term care patients (439) - SCI (330) - Cancer (203)
Refs [20,28,31]	Refs [19,20,27,28]	
		- Hypertension and age 75-84 years (878)
- Comparing HIV+ patients with and without AIDS - Survivors with non-survivors: 5.47 vs. 7.20		- SCI (330) - HIV + (395) - ICU (105) - Breast cancer (>10,000) - Asthma (>10,000)

APPENDIX - Continued

Content validity	Concurrent validity	Predictive validity	Construc
[Disease count - ctnd]			
Information needed:			
- interview	- APACHE II = 0.53	- physical limitations: r^2 = 0.38	Correlat
- patient record	- ASA score = 0.20	- ADL: r^2 = 0.14	- Age = 0
- ICD-9 codes		- IADL: r^2 = 0.17	- LOS = 0
Final score:			- Charge
- number of present diseases, maximal score depends on the limits that are applied			- inpatie - nosoco
Adaptations:			
- several			
Refs [19,33-36]	Refs [19,36]	Refs [33,35]	Refs [33,3
DUSOI index			
Items:	Correlation with:	Multivariate regression models	Correlat
Every present health problem is rated on four domains:	- VAS severity rating: ICC = 0.61	predicting:	- Age = 0
- symptom level;	- Patient record vs clinician scores:	- number of follow-up visits	- Gender
- complication level;	r = 0.53-0.67	- >6 visits	- Employ
- prognosis without treatment;		- referral	- Educat
- prognosis with treatment		- follow-up severity >37.5	
Weights:		- costs	
- 0-5		- OR = 1.02-1.03	
		- AUC = 0.60-0.69 (monovariate)	
		- AUC = 0.65-0.72 (multivariate)	
Information needed:			
- clinical judgment			
Final score:			
- scoring paradigm to a 0-100 score			
Refs [37]	Refs [37]	Refs [38]	Refs [37]
Hallstrom index			
Items:		Logistic regression model predicting:	Mean sc
Two domains:		Mortality	non-surv
- CF consisting of 10 conditions;		- OR = 1.51	0.87 vs.
		- r^2 increase from 9 to 11.2%	
- SF consisting of 6 cardiac symptoms			Correlat
			Age = 0.

Construct validity	Reliability	Study populations (n)
Correlation with: - Age = 0.22 - LOS = 0.27 - Charges = 0.19 - inpatient mortality - nosocomial infections Refs [33,34,36]		- Myocardial infarction (>10,000) - Appendicitis (>10,000) - Abdominal hernia (>10,000) - Diverticulitis (>10,000) - Biliairy tract disease (>10,000) - Low back pain (>10,000) - Pneumonia (>10,000) - Diabetes with complications (>10,000) - Rehabilitation inpatients (>10,000)
Correlation with: - Age = 0.61 - Gender = 0.12 - Employment = -0.20 - Education = -0.15 Refs [37]	Test-retest: - r = 0.59, 0.65 Intra-rater: - ICC = 0.39-0.78 - r = 0.67, 0.76 Inter-rater: - ICC = 0.45 - r = 0.47-0.63 Refs [37-39]	- Mixed primary care patients (414+1191)
Mean scores of survivors vs. non-survivors: 0.87 vs. 1.08, $p < 0.0005$ Correlation with: Age = 0.21		- Out of hospital ventricular fibrillation (282)

APPENDIX - Continued

Content validity	Concurrent validity	Predictive validity	Construct
[Hallstrom index - ctnd]			
Weights:			
- none			
Information needed:			
- interview			
Final score:			
- CF: number of present conditions;			
- SF: number of present symptoms.			
- Total 1.67 x CF + SF			
Adaptations:			
- none			
Refs [40]		Refs [40]	Refs [40]
Hurwitz index			
Items:		Regression model predicting	
- no comorbidity		whether patients receive	
- non-disabling comorbidity		medical or chiropractic care:	
- disabling comorbidity		- non-disabling comorbidity: OR = 0.69	
Information needed:		- disabling comorbidity: OR = 0.37	
- not specified			
Adaptations:			
- none			
Refs [41]		Refs [41]	
ICED			
Disease Severity (DS)	Correlation with:	Multivariate regression models	
Items:	- ASA = 0.53	predicting:	
(194)			
- 14 disease categories	- BOD = 0.55	- serious complications	
Weights:	- Charlson Index = 0.69	- IADL	
- 1-5			
Information needed:		- score 2: OR = 2.4	
- symptoms, signs and		- score 3: OR = 4.5	
laboratory tests		- score 4: OR = 16.8	

Construct validity	Reliability	Study populations (n)
Refs [40]		- Back-related problems (931)
	Intra-rater: - ICED: ICC = 0.95 - DS: ICC = 0.99 - FS: ICC = 0.81 Inter-rater: - ICED = 0.57, 0.71 - DS = 0.60, 0.71 - FS = 0.75, 0.77	- Total hip replacement (356) - Long-stay nursing home residents

APPENDIX - Continued

Content validity	Concurrent validity	Predictive validity	Constru
[ICED - ctnd]			
Functional Severity (FS)			
Items:			
- 10 functional areas			
Weights:			
- 1-3			
Information needed:			
- level of impairment			
Final score:			
- scoring paradigm to a 1-4 score			
Refs [42]	Refs [4,21,42]	Refs [42]	
Incalzi index			
Items:		Multivariate regression model predicting:	
- 52 conditions		Mortality	
Weights:		- dichotomised index with	
- based on RR for mortality		cut-off 5: OR = 1.58	
Information needed:		- dichotomised age- index with	
- history		cut-off 7: OR = 1.77	
- physical examination			
- laboratory data			
- ECG		Comparing mortality for patients	
- chest X- ray		with scores >75th percentile with	
Final score:		those with scores <75th percentile:	
- sum of weights		- 20 vs. 5.3%, p < 0.001	
Adaptations:			
- adjustment for age >75			
Refs [44]		Refs [44]	
Kaplan index			
Items:		Type:	
- vascular or non-vascular disease		- non-vascular: 5-yr fatality rate = 40%	
Weights:		- vascular: 5-yr fatality rate = 58%	
0 – non cogent, easy to control or no comorbidity;		Severity level:	
1 – slight decompensation of vital system or non-threatening		0: 5-yr fatality rate = 7%	
		1: 5-yr fatality rate = 28%	

Construct validity	Reliability	Study populations (n)
	Refs [43]	
	Inter-rater: - 'good'	- Mixed geriatric and general medicine (370)
	Refs [44]	
%		- Diabetes (188) - Breast cancer (404)

APPENDIX - Continued

Content validity	Concurrent validity	Predictive validity	Constru
[Kaplan index - ctnd]			
chronic conditions;		2: 5-yr fatality rate = 42%	
2 – impaired vital system or potentially threatening chronic condition;		3: 5-yr fatality rate = 69%	
3 – recent full decompensation of vital system or life-threatening chronic condition		Multivariate regression model predicting:	
Information needed:		- 10-yr mortality: RR = 1.98	
- clinical information		- 63-month mortality: RR = 1.08-1.15	
Final score:			
- according to the most severe condition, two grades 2 are ranked as 3			
Adaptations:			
- MMCI, ACE-27			
Refs [45- 47]		Refs [5,14,47]	
Liu index			
Items:	Correlation with:	Multivariate regression model predicting:	Correla
- 38 conditions	- Charlson Index = 0.40	- FIM: r^2 increase from 0.73 to 0.80	- numb
Weights:		- LOS: r^2 increase from 0.37 to 0.45	- labora
0 (not present) to 5 (active rehabilitation contra-indicated)			- interv - consu - interr
Information needed:			- age
- medical records			- LOS
Final score:			- admis
- sum of weights			- discha
Adaptations:			range fr
- None			
Refs [18]	Refs [18]	Refs [18]	Refs [18,
Shwartz index			
Items:		Multivariate regression model predicting costs using Shwarz index based on:	
- 21 conditions			
Weights:		- medical records: r^2 increase from 0.42 to 0.50	
- regression coefficient from a model to predict costs			

	Construct validity	Reliability	Study populations (n)
ting:	Correlations with: - number of medications - laboratory studies - interventions - consultations - interrupted days - age - LOS - admission FIM - discharge FIM range from 0.22-0.61	Inter-rater: - ICC = 0.997	- Stroke (106)
	Refs [18,48]	Refs [18,48]	
		Intra-rater: - 'high'	- Mixed patients (4439) - Stroke - Lung disease - Heart disease - Prostate disease

APPENDIX - Continued

Content validity	Concurrent validity	Predictive validity	Construc
[Shwartz index - ctnd]			
Information needed:		- ICD-9 codes: r^2 increase	
- medical records or databases with ICD-9 codes		from 0.42 to 0.44	
Final score:			
- sum of weights of present conditions			
Refs [49]		Refs [49]	

ACE-27 = Adult Comorbidity Evaluation 27 items; ADL = Activities of Daily Living; APACHE II = Acute Physiologic and Chronic Health Evaluation II scoring system; ASA = American Society of Anaesthesiologists score; AUC = Area under the Receiver Operating Characteristic Curve; BOD = Burden of Disease index; CF = Chronic Factor; CIRS = Cumulative Illness Rating Scale; DUSOI = Duke Severity of Illness; ECG = Electro Cardio Gram; FIM = Functional Independence Measure; HIV+ = Human Immunodeficiency Virus positive; IADL = Instrumental Activities of Daily Living; ICC = Intra Class Correlation Coefficient; ICD-9 = International Classification of

Construct validity	Reliability	Study populations (n)
		- Hip and femur fracture - Low back disorder
	Refs [49]	

Diseases, version 9; ICED = Index of Coexistent Disease; ICU = Intensive Care Unit; K = Kappa; LOS = Length of Stay; LT = Long term; MMCI = Modified Medical Comorbidity Index; OR = Odds Ratio; r = Pearson or Spearman correlation coefficient; r^2 = explained variance; Refs = references; RR = Relative Risk; SCI = Spinal Cord Injury; SF = Symptom Factor; SIP = Sickness Impact Profile; ST = Short term; VAS = Visual Analogue Scale; Weights = description of method used to apply weights to comorbid conditions; wK = Weighted Kappa.

2a

How to measure comorbidity: a critical review of available methods (Authors reply)

Vincent de Groot, Heleen Beckerman,
Gustaaf J. Lankhorst & Lex M. Bouter

J Clin Epidemiol 2004; 57:323

We would like to thank Rozzini *et al.* for their letter regarding the Geriatric Index of Comorbidity (GIC). Unfortunately the GIC was not included in our review, because it was only recently published.[1] In this letter we will briefly review the GIC on the basis of their publication and letter to the editor according to the criteria we used in our review of available comorbidity measures.

The selection of the 15 diseases is based on the prevalence of these conditions in an elderly population. This selection and the severity rating of the condition are based on the Index of Coexistent Disease – Disease Severity (ICED-DS).[2,3] Subsequently, the GIC class is determined. Although the authors give a clear description of their classification, they do not describe the scientific basis for it. Medical charts are needed to collect the information. Their study population consisted of 493 elderly patients admitted to the Geriatric Evaluation and Rehabilitation Unit[1] and 1402 hospitalised elderly patients (letter). Concurrent validity was studied by correlating the GIC with the number of comorbid diseases (selected from the 15 diseases and stratified into four levels) and the sum of severities of the 15 conditions (stratified into four levels). The correlation coefficients were 0.14 and 0.53, respectively.[1] Unfortunately, the authors do not describe why they decided to use stratification. Predictive validity was assessed by using the GIC in multivariate models predicting Basic ADL (BADL; r^2=0.32) and Physical Performance Test (PPT; r^2=0.39). A relative risk (RR) of 2.3 (95%CI 1.7-3.1) was found for the GIC in a multivariate model predicting twelve month mortality.[1] An RR of 1.1 (95%CI 1.0-2.1) was found for Class III and an RR of 3.0 (95%CI 1.7-5.3) for Class IV was found in a multivariate Cox regression model predicting 6-month mortality (letter). The relationships between the GIC and BADL, PPT, Minimal Mental State Examination, Acute Physiological And Chronic Health Evaluation II (APACHE II) scores and length of stay support construct validity. Inter-rater agreement is 89% and intra-rater agreement is 97%.

Interestingly, the authors showed that the GIC was independently associated with BADL and PPT after adjustment for severity

of individual diseases. Thus, the GIC seems to contain some information that is not captured by using the number of diseases and their severity ratings only. In our review we discussed that certain disease combinations may have synergistic effects, leading to more disability than would be expected on the basis of addition. Knowing these combinations provides extra information. It would be very interesting to study these phenomena in depth.

Although the GIC certainly seems promising, further validation studies are needed. We are very interested in studies comparing the GIC with other comorbidity indices mentioned in our review and more extensive reliability studies.

References

1. Rozzini R, Frisoni GB, Ferrucci L, Barbisoni P, Sabatini T, Ranieri P *et al.* Geriatric Index of Comorbidity: validation and comparison with other measures of comorbidity. Age Ageing 2002; 31:277-285.
2. Greenfield S, Apolone G, McNeil BJ, Cleary PD. The importance of co-existent disease in the occurrence of postoperative complications and one-year recovery in patients undergoing total hip replacement. Comorbidity and outcomes after hip replacement. Med Care 1993; 31:141-154.
3. de Groot V, Beckerman H, Lankhorst GJ, Bouter LM. How to measure comorbidity. a critical review of available methods. J Clin Epidemiol 2003; 56:221-229.

3

The usefulness of evaluative outcome measures in patients with multiple sclerosis

Vincent de Groot, Heleen Beckerman, Bernard M.J. Uitdehaag,
Riekie C.W. de Vet, Gustaaf J. Lankhorst,
Chris H. Polman & Lex M. Bouter

Brain 2006; 129:2648-2659

Abstract

Objectives. To select the most useful evaluative outcome measures for multiple sclerosis (MS).

Methods. We included 156 recently diagnosed patients in a 3-yr follow-up study, and assessed them on 23 outcome measures in the domains of disease-specific outcomes, physical functioning, mental health, social functioning and general health. A Global Rating Scale (GRS) and the Expanded Disability Status Scale (EDSS) were used as external criteria to determine the Minimally Important Change (MIC) for each outcome measure. Subsequently, we determined whether the outcome measures could detect their MIC reliably. From these, per domain the outcome measure that was found to be most sensitive to changes (responsive) was identified.

Results. At group level, 11 outcomes of the domains of physical functioning, mental health, social functioning and general health could reliably detect the MIC. Of these 11, the most responsive measures per domain were the Medical Outcome Study 36 Short Form sub-scale physical functioning (SF36pf), the Disability and Impact Profile (DIP) sub-scale psychological, the Rehabilitation Activities Profile sub-scale occupation (RAPocc) and the DIP sub-scale mental health, respectively. Overall, the most responsive measures were the SF36pf and the RAPocc. In individual patients, none of the measures could reliably detect the MIC.

Conclusions. In the early stages of MS the most useful evaluative outcome measures for research are the SF36pf (physical functioning) and the RAPocc (social functioning).

Introduction

The Expanded Disability Status Scale (EDSS) is a frequently used and well-known outcome measure for multiple sclerosis (MS). However, it is criticized because it has unsatisfactory validity, and its reliability is poor.[1-3] In response to this situation, the National MS Society Clinical Outcomes Assessment Task Force reviewed a large number of data sets to determine which outcome measures would adequately reflect the consequences of MS and are capable of reliably assessing these consequences.[4,5] This led to the development of the MS Functional Composite Measure (MSFC), which consists of the timed 25-foot walk test (TWT), the Nine-Hole Peg Test (NHPT) and the Paced Auditory Serial Addition Test (PASAT). Originally, the Task Force intended to include a measure of visual acuity, but no reliable measure could be found. The MSFC is intended to replace the EDSS as outcome measure in current and future trials.[4,6,7] The interpretation of the scores of the individual components of the MSFC is straightforward. However, the total score, which results from a relatively complex formula to combine the component scores, is more difficult to interpret. An adaptation of the MSFC, the Short and Graphic Assessment Scale (SaGAS),[8] uses only the TWT and the NHPT. Through specific transformation, a score is obtained that should be easier to interpret. Other newly developed disease-specific outcomes are the Multiple Sclerosis Impact Scale,[9] and the Guy's Neurological Disability Scale.[2] In addition to these new, disease-specific, measures, several other Disability and Quality of Life measures have been used in MS research.[10-19]

Responsiveness is an important clinimetric property. It represents the ability to measure change, and is particularly relevant when outcome measures are to be used in longitudinal studies, such as clinical trials.[20,21] In connection with MS, however, it has been studied much less extensively than validity and reliability.[2,22-29] Moreover, in the literature there is no consensus about the exact definition of responsiveness.[20] Consequently, there are many cur-

rently available methods that have been developed to assess responsiveness.[20,30,31] It has been shown that applying different methods leads to different conclusions about the *absolute* responsiveness of an outcome measure.[20] However, conclusions about the *relative* responsiveness, *i.e.* how do different measures perform in relation to each other, are less dependent on the method used.[20] To assess the *relative* responsiveness, several outcome measures of interest should be included, and parallel assessments should be made at the same points in time.

The methods that can be used to assess whether scores have changed can be sub-divided into distribution-based and anchor-based methods.[32-35] Distribution-based methods, using standardized metrics, focus on the ability of an outcome measure to reliably determine change, and aim to quantify the noise, *i.e.* the variability of the score changes in the absence of a relevant change. Anchor-based methods focus on the correspondence of the change on the outcome measure of interest with the change on an external criterion[32,36] and aim to quantify the signal, *i.e.* the size of the score change when there is a relevant change. The results of anchor-based methods depend on the external criterion and the cut-off point chosen.[32] The usefulness of an evaluative outcome measure depends on whether score changes associated with a relevant change can reliably be distinguished from the variability of score changes in absence of a relevant change.[37]

In this study, 23 (sub-scales of) outcome measures were compared. The aim was to select the most useful evaluative outcome measures for the early stages of MS.

Methods

Patients

All consecutive potentially eligible patients visiting the participating neurology outpatient clinics were invited to participate. A cohort of 156 recently (< six months previously) diagnosed patients, aged 16-55 years, was recruited and followed prospective-

ly for three years. Diagnosis was based on the Poser-criteria for definite MS.[38] Patients with other neurological disorders, or systemic or malignant neoplastic diseases, were excluded. The measurements took place at baseline, and six months, and after one, two and three years. In the case of a relapse, the measurements were postponed for a few weeks until the relapse had subsided. The patients were visited at home in order to minimize drop-out. Four well-trained raters were responsible for the scoring.

Outcome measures

We studied the (sub-)scales of the EDSS,[39-41] the MSFC,[4-6,42-45] the SaGAS,[8] the Action Research Arm Test (ARAT),[46,47] the Disability and Impact Profile (DIP),[15-18,48] the Functional Independence Measure (FIM),[10,12,49] the Rehabilitation Activities Profile (RAP),[50,51] the Rivermead Mobility Index (RMI),[52-55] and the Medical Outcome Study Short Form 36 (SF36).[11,14,56,57] The 23 (sub-)scales covered five domains: three disease-specific measures, ten physical functioning measures (five mobility measures, three self care measures, and two upper limb function measures), four mental health measures (two cognitive function measures and two emotional well-being measures), five social functioning measures and one general health measure. Of these, 11 outcome measures were questionnaires, seven were (parts of) measures that required physical examination or testing procedures, and five outcome measures were based on semi-structured interviews. When possible, outcome measures were transformed into a scale ranging from 100 (best) to 0 (worst). Scores on the NHPT, the 10-meter Timed Walk Test (TWT), the MSFC, and the SaGAS could not be transformed in this way, because these continuous scales do not have defined end-points for best or worst scores. TABLE 3.1 presents an overview of the outcome measures and the baseline scores (standard deviation).

Analysis of responsiveness

To determine whether a patient's score had changed, we applied

TABLE 3.1 Outcome measures studied and baseline scores of 156 MS patients.

Outcome measure		Sub-scale	Type	Transformed baseline score [0-100% (SD)]
Disease-specific				
EDSS	Expanded Disability Status Scale		pt	74.9 (11.2)
MSFC	Multiple Sclerosis Functional Composite		pt	0.0 (0.7)*
SaGAS	Short and Graphic Assessment Scale		pt	7.0 (0.4)*
Physical functioning				
Mobility				
DIPmob	Disability and Impact Profile	Mobility	q	86.9 (10.5)
RAPmob	Rehabilitation Activities Profile	Mobility	i	85.7 (14.1)
RMI	Rivermead Mobility Index		q	95.7 (8.7)
SF36pf	Medical Outcome Study Short Form 36	Physical Functioning	q	71.3 (23.5)
TWT	10-meter Timed Walk Test		pt	6.4 (3.2) s*
Self care				
DIPself	Disability and Impact Profile	Self-care	q	94.3 (8.6)
FIMmf	Functional Independence Measure	Motor Function	i	95.2 (5.4)
RAPself	Rehabilitation Activities Profile	Self-care	i	92.3 (11.1)
Upper limb function				
ARAT	Action Research Arm Test		pt	99.1 (4.0)
NHPT	Nine Hole Peg Test		pt	21.1 (4.0) s*
Mental health				
Cognitive function				
FIMcf	Functional Independence Measure	Cognitive Function	i	95.2 (5.2)
PASAT3	Paced Serial Addition Test	3-second version	pt	76.9 (18.3)*
Emotional well-being				
DIPpsy	Disability and Impact Profile	Psychological	q	79.4 (12.3)
SF36mh	Medical Outcome Study Short Form 36	Mental Health	q	72.1 (17.5)
Social functioning				
DIPsoc	Disability and Impact Profile	Social Functioning	q	87.0 (10.2)
RAPocc	Rehabilitation Activities Profile	Occupation	i	75.0 (20.6)
SF36re	Medical Outcome Study Short Form 36	Role Emotional	q	74.1 (37.0)
SF36rp	Medical Outcome Study Short Form 36	Role Physical	q	51.9 (42.0)
SF36sf	Medical Outcome Study Short Form 36	Social Functioning	q	77.9 (23.2)
General health				
SF36gh	Medical Outcome Study Short Form 36	General Health	q	52.6 (19.8)

pt = performance test; q = questionnaire; i = interview by professional. *Not transformed into a 100 (best) to 0 (worst) scale

two external criteria: (i) a 7-point Likert-type patient rated Global Rating Scale (GRS) of change, using the situation at diagnosis as reference point,[33,58-63] emphasizing the perspective of the patient, and (ii) a change on the EDSS, representing the perspective of the clinician. The GRS question asked was: 'How would you rate your current health when compared with your health at the time of diagnosis?' The answering categories were: very much improved, much improved, slightly improved, stable, slightly deteriorated, much deteriorated, and very much deteriorated. The EDSS is a single-scale measure that ranges from 0 = a normal neurological examination, to 10 = death due to MS.

To assess the *relative* responsiveness, that is relatively independent of the method used to assess the responsiveness,[20] we calculated the Area Under the Receiver Operating Characteristic (ROC) curve with its 95% Confidence Interval (AUC, 95%CI) for every outcome measure, using score changes since baseline at three years.[21,64-66] We used a non-parametric method which does not make any assumptions about the distributions to compute the AUC. FIGURE 3.1 shows an example of two ROC curves. The *relative* responsiveness was assessed separately for deterioration and improvement. For both external criteria the scores were

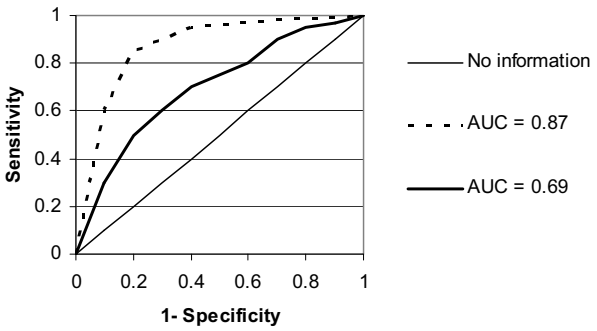


FIGURE 3.1 ROC curves. In a ROC curve the sensitivity is plotted against 1- specificity. The AUC is a measure of the responsiveness of the outcome measure. An AUC ≤0.5 (diagonal line) indicates that the outcome measure is not responsive. The more the ROC curve approaches the upper left corner the more responsive the outcome measure is.

dichotomized, using the category stable (no change) as reference category.

The Minimally Important Change score of an outcome measure (MIC) is calculated as the mean change score in patients who showed minimally important change according to an external criterion.[67] For the GRS of the patient's perspective we used the categories of slightly improved or slightly deteriorated to identify the patients who showed a minimally important change. FIGURE 3.2 illustrates graphically were the MIC is located on the spectrum of

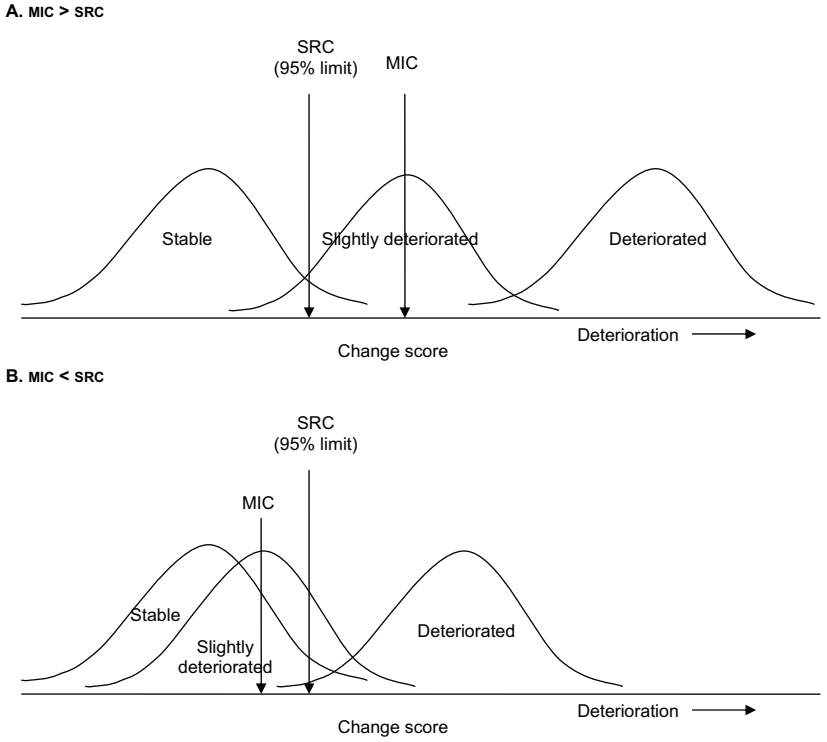


FIGURE 3.2 Relationship between SRC and MIC. (A) shows the distribution of change scores for the categories (stable, slightly deteriorated and deteriorated) of the external criterion. There is minimal overlap between scores and the MIC is much larger than the SRC. This outcome measure is useful. (B) Shows again the distribution of change scores for each category of the external criterion. There is much overlap between the scores and the MIC is smaller than the SRC. This outcome measure is not useful.

change scores. The next possible categories, namely much improved or much deteriorated, were not used, because they indicate substantial improvement or deterioration. For EDSS of the clinician's perspective we used an improvement or deterioration of one point since baseline, because a change of one EDSS point is frequently used in trials and is the lowest EDSS change that can reliably be detected in the lower EDSS ranges.[68,69] The MIC was calculated from the patient's perspective (MIC-$P_{improvement}$ and MIC-$P_{deterioration}$), and the clinician's perspective (MIC-$C_{improvement}$ and MIC-$C_{deterioration}$). Because the longitudinal study design had five repeated measurements, we used Generalised Estimating Equations (GEE) to estimate the MIC. This regression analysis technique for longitudinal data makes optimal use of the available data and reduces the standard error of the estimates, while at the same time correcting for the dependence between subsequent measurements.[70] The correlation structure was chosen on the basis of the correlation matrix of the outcome measures, and set at exchangeable (*i.e.* correlation coefficients between the first and successive measurements are approximately equal) for all outcomes except the cognitive subscale of the FIM that was set at 4-dependence (*i.e.* correlation coefficients between the first and successive measurements are progressively smaller). Scores on the outcome measures were used as dependent variable [Y(t)], and time (t, in years) and four dummy variables based on the external criteria (deteriorated, slightly deteriorated, slightly improved, improved) were used as independent variables. The stable group was used as reference. Because the GRS used the time of diagnosis as reference point, we used an autoregression formula that also includes the score for the outcome measure at baseline [$Y(t_0)$] as independent variable. In formula:

$$Y(t) = \alpha + \beta_1 \times Y(t_0) + \beta_2 \times t + \beta_3 \times \text{deteriorated} + \beta_4 \times \text{slightly deteriorated} + \beta_5 \times \text{slightly improved} + \beta_6 \times \text{improved}$$

β_4 is interpreted as the mean score change on the outcome measure for patients who were slightly deteriorated, and provides an

estimate for the $MIC_{deterioration}$. β_5 is interpreted as the mean score change on the outcome measure for patients who were slightly improved, and provides an estimate for the $MIC_{improvement}$.

To assess the reliability of two scores on each outcome measure, we used the Smallest Real Change (SRC).[21,71,72] The SRC is more often referred to as the Smallest Real Difference, but since our main focus is on intra-individual *changes*, we prefer to use the term Smallest Real Change. For each external criterion the SRC was calculated in the sub-group of patients who did not change, according to the external criterion during the first six months after inclusion. The SRC takes two sources of variability into account: (i) the reliability of the outcome measure, and (ii) the naturally occurring variability in stable patients. The SRC offers the opportunity to calculate a measure for comparisons at group level (SRC_{group}) and at individual level ($SRC_{individual}$).[71] The $SRC_{individual}$ was calculated as 1.96 x SD of the score changes in stable patients. FIGURE 3.2 shows graphically were the SRC is located on the spectrum of change scores. The SRC_{group} was calculated as $SRC_{individual} / \sqrt{n}$.

The selection of the most useful evaluative outcome measure was based on the *relative* responsiveness (highest AUC), whether the MIC > $SRC_{individual}$ or SRC_{group} (*see* Fig. 3.2), and whether the results were comparable for both external criteria. For each outcome measure we calculated the sample sizes (patients per group) needed to show differences between independent samples in future studies. We used the formula $2 \times \{[(Z_\alpha + Z_\beta) \times (SRC_{group} / 1.96)] / MIC\}^2$,[37] where α is set at 0.05 ($Z_\alpha = 1.96$) and β is set at 0.20 ($Z_\beta = 0.84$), in order to achieve a power of 0.80.

The statistical analyses were performed with SPSS version 11.5 for Windows. GEE analyses were performed with the Statistical Package for Interactive Data Analysis (SPIDA) version 6.05 from the Statistical Computing Laboratory.

Results

A total of 156 patients were included in the cohort between January

TABLE 3.2 Baseline characteristics of patients with multiple sclerosis.

	Relapse onset		Non-relapse onset		
	RR	SP	PP	Not yet known	Total
n (%)	120 (77%)	8 (5%)	25 (16%)	3 (2%)	156 (100%)
Age (SD)	35.5 (8.9)	48.2 (6.7)	43.2 (8.9)	45.5 (6.9)	37.6 (9.5)
Female (%)	84 (70.0%)	3 (37.5%)	11 (44%)	3 (100%)	101 (64%)
Years since diagnosis	0.26 (0.15-0.41)	0.33 (0.24-0.48)	0.28 (0.15-0.33)	0.14 (0.14-0.17)	0.26 (0.15-0.40)
Years since symptoms	1.83 (0.67-4.40)	7.50 (3.35-14.51)	2.10 (1.07-3.15)	3.62 (3.53-4.63)	2.15 (0.79-4.36)
Number of exacerbations	2 (1-3)	2 (1-7)	0 (0-0)	0 (0-0)	2 (1-3)
EDSS	2.0 (2.0-3.0)	3.0 (2.5-3.9)	3.0 (2.5-4.0)	2.5 (2.0-4.0)	2.5 (2.0-3.0)

n (percentage), mean (SD) or median (IQR). RR = relapsing remitting multiple sclerosis; SP = secondary progressive multiple sclerosis; PP = primary progressive multiple sclerosis. EDSS = Expanded Disability Status Scale, original score.

1998 and January 2001. TABLE 3.2 shows the baseline characteristics of these patients. Most characteristics comply with the expected pattern: more females than males in the Relapsing-Remitting group, more males than females in the Primary Progressive group, and more severe neurological deficits in the Primary Progressive group. Seven patients were lost to follow-up (three after one year, one after two years and three after three years), and 15 measurements were missing. The baseline scores on the outcome measure are presented in TABLE 3.1.

TABLE 3.3 shows the distribution of GRS and EDSS scores for each measurement. The distributions are remarkably different. The GRS scores are more equally spread across the categories, and according to the GRS less patients were stable, and more patients had improved. Over time there is a tendency for both external criteria to change towards deterioration. The percentage of patients that deteriorated (taking categories deteriorated and slightly deteriorated together) according to the patient's and clinician's perspective, respectively, is 36 and 22% at six months, 46 and 33% at one, 50 and 46% at two, and 60 and 44% at three years. The agreement between the patient's and clinician's perspective to classify

TABLE 3.3 Distribution (n, %) of the GRS (patient's perspective) and EDSS (clinician's perspective) based external criteria for each measurement.

	Patient's perspective (%)				Clinician's perspective (%)			
	6 months (n = 113)	1 year (n = 130)	2 years (n = 141)	3 years (n = 145)	6 months (n = 153)	1 year (n = 147)	2 years (n = 145)	3 years (n = 146)
Deteriorated	11 (10)	15 (12)	19 (13)	28 (19)	12 (8)	24 (16)	40 (28)	41 (28)
Slightly deteriorated	29 (26)	44 (34)	52 (37)	60 (41)	21 (14)	25 (17)	26 (18)	24 (16)
Stable	26 (23)	30 (23)	29 (21)	22 (15)	100 (65)	79 (54)	66 (46)	69 (47)
Slightly improved	14 (12)	11 (8)	19 (13)	10 (7)	11 (7)	11 (7)	11 (8)	8 (5)
Improved	33 (29)	30 (23)	22 (16)	25 (17)	9 (6)	8 (5)	2 (1)	4 (3)

GRS = Global Rating Scale; EDSS = Expanded Disability Status Scale. For deterioration and improvement the categories 'very much' and 'much' have been combined.

patients as deteriorated, stable or improved is 35% (kappa 0.10) at six months, 42% (kappa 0.14) at one, 40% (kappa 0.07) at two, and 45% (kappa 0.13) at three years.

TABLES 3.4 and 3.5 show that the AUCs range from 0.50 to 0.75 and have wide confidence intervals. For five (patient's perspective) and seven (clinician's perspective) outcome measures the AUC does not significantly differ from 0.50. For a substantial number of outcome measures the MIC does not significantly differ from zero, which means that the MIC cannot be detected beyond chance for these outcome measures in this population. It also means that these outcome measures are not suitable to evaluate change in this population. Furthermore, none of the outcome measures has an MIC > $SRC_{individual}$, which makes the outcome measures unsuitable to detect a minimally important change in an individual patient. However, several measures have an MIC > SRC_{group}, which makes them suitable for research purposes. The final columns in the tables show a large variation in required sample sizes. The unrealistically high estimates of the sample sizes are caused by large estimates of the $SRC_{individual}$ relative to the estimate of the MIC.

The results for *deterioration* from the *patient's* perspective can be found in TABLE 3.4. Of the disease-specific outcome measures, the

EDSS has the highest AUC [0.70 (95%CI 0.62-0.79)]. For all three disease-specific outcome measures the MIC-$P_{deterioration}$ is small, and does not significantly differ from zero. Of the outcome measures related to physical functioning, the SF36 sub-scale physical functioning (SF36pf) has the highest AUC [0.75 (95%CI 0.67-0.84)] and a MIC-$P_{deterioration}$ (-8.58) that exceeds the SRC_{group} (-4.38). Of the outcome measures related to mental health, the FIM sub-scale cognitive function (FIMcf) and the DIP sub-scale psychological (DIPpsy) have approximately the same AUCs [0.65 (95%CI 0.55-0.74) and 0.64 (95%CI 0.55-0.73), respectively]. For the DIPpsy the MIC-$P_{deterioration}$ (-2.88) exceeds the SRC_{group} (-2.80), but for the FIMcf the MIC-$P_{deterioration}$ (-1.47) is smaller than the SRC_{group} (-1.66). Of the outcome measures related to social functioning, the RAP sub-scale occupation (RAPocc) has the highest AUC [0.73 (95%CI 0.64-0.81)] and a MIC-$P_{deterioration}$ (-7.74) exceeding the SRC_{group} (-4.24).

TABLE 3.5 shows the results for *deterioration* from the *clinician's* perspective. Because information from the EDSS is used to obtain the external criterion, results for the EDSS cannot be calculated. The two disease-specific outcome measures have a very similar AUC [0.72 (95%CI 0.63-0.81) for the SaGAS and 0.71 (95%CI 0.62-0.80) for the MSFC], and for both the MIC-$C_{deterioration}$ was small and did not significantly differ from zero. Of the outcome measures related to physical functioning, the SF36pf has the highest AUC [0.72 (95%CI 0.63-0.80)] and a MIC-$C_{deterioration}$ (-8.52) that amply exceeds the SRC_{group} (-2.81). Of the outcome measures related to mental health, the DIP sub-scale psychological function (DIPpsy) and the Paced Serial Addition Test 3 second version (PASAT3) have an AUC of 0.60 (95%CI 0.50-0.70 and 0.50-0.69, respectively). For both outcome measures the MIC-$C_{deterioration}$ is small and does not significantly differ from zero. Of the outcome measures related to social functioning, the RAPocc has the highest AUC [0.69 (95%CI 0.61-0.78)] and a MIC-$C_{deterioration}$ (-8.40) that amply exceeds the SRC_{group} (-2.69).

Regardless of the domain of the outcome measures, the five most responsive (AUC) outcome measures to detect *deterioration*

TABLE 3.4 AUC, MIC-P, and SRC for deterioration using the patient's perspective as external criterion.

	AUC	95%CI	$MIC_{deterioration}$	$SRC_{individual}$	SRC_{group}	Sample Size*
Disease-specific						
EDSS	0.70	0.62-0.79	-1.50 (ns)	-16.04	-3.15	467
MSFC**	0.62	0.53-0.72	-0.05 (ns)	-0.54	-0.11	476
SaGAS**	0.65	0.56-0.75	-0.05 (ns)	-0.25	-0.05	102
Physical functioning						
Mobility						
DIPmob	0.73	0.65-0.82	-4.25	-8.99	-1.80	18
RAPmob	0.66	0.57-0.76	-3.42	-19.88	-3.90	138
RMI	0.67	0.58-0.76	-0.88 (ns)	-5.91	-1.16	184
SF36pf	0.75	0.67-0.84	-8.58	-21.91	-4.38	27
TWT**	0.65	0.56-0.74	1.15 (ns)	2.56	0.50	20
Self care						
DIPself	0.70	0.62-0.79	-2.11	-9.54	-1.91	83
FIMmf	0.68	0.59-0.76	-1.45	-5.74	-1.13	64
RAPself	0.65	0.56-0.74	-2.41	-11.96	-2.35	101
Upper limb function						
ARAT	0.53	0.43-0.63	-0.06 (ns)	-1.61	-0.32	2939
NHPT**	0.59	0.49-0.69	0.30 (ns)	2.82	0.55	361
Mental health						
Cognitive function						
FIMcf	0.65	0.55-0.74	-1.47	-8.47	-1.66	136
PASAT3	0.50	0.40-0.60	2.56	19.62	4.18	240
Emotional well-being						
DIPpsy	0.64	0.55-0.73	-2.88	-14.01	-2.80	97
SF36mh	0.56	0.46-0.66	-4.45	-28.13	-5.63	163
Social functioning						
DIPsoc	0.68	0.59-0.77	-2.84	-8.08	-1.62	33
RAPocc	0.73	0.64-0.81	-7.74	-21.63	-4.24	32
SF36re	0.50	0.40-0.59	-8.13	-67.26	-13.45	279
SF36rp	0.60	0.51-0.69	-21.69	-92.24	-18.45	74
SF36sf	0.68	0.59-0.77	-11.15	-41.17	-8.23	56
General health						
SF36gh	0.66	0.57-0.75	-9.86	-26.61	-5.32	30

AUC = area under the ROC curve at three years after baseline with 95%CIs; $MIC_{deterioration}$ = Minimally Important Change; $SRC_{individual}$ = Smallest Real Change at individual level; SRC_{group} = Smallest Real Change at group level, based on 26 stable patients at six months; GRS = Global Rating Scale; ns = not significantly different from 0. * Patients per group, calculation based on a significance level of 0.05 and a power of 0.8. ** MSFC, SaGAS, NHPT, TWT data not transformed into a 0-100 point scale.

TABLE 3.5 AUC, MIC-P, and SRC for deterioration using the clinician's perspective as external criterion.

	AUC	95%CI	$MIC_{deterioration}$	$SRC_{individual}$	SRC_{group}	Sample Size*
Disease-specific						
EDSS						
MSFC**	0.71	0.62-0.80	0.08 (ns)	-0.72	-0.08	331
SaGAS**	0.72	0.63-0.81	-0.06 (ns)	-0.44	-0.04	220
Physical functioning						
Mobility						
DIPmob	0.66	0.57-0.75	-2.56	-10.52	-1.06	69
RAPmob	0.67	0.58-0.76	-5.62	-16.26	-1.63	34
RMI	0.65	0.56-0.75	-1.30	-7.53	-0.75	137
SF36pf	0.72	0.63-0.80	-8.52	-27.99	-2.81	44
TWT**	0.69	0.59-0.78	0.34 (ns)	3.03	0.30	324
Self care						
DIPself	0.65	0.55-0.74	-2.16	-8.70	-0.87	66
FIMmf	0.68	0.59-0.77	-1.70	-6.43	-0.64	58
RAPself	0.62	0.52-0.72	-1.33 (ns)	-14.79	-1.48	505
Upper limb function						
ARAT	0.55	0.45-0.65	-0.14 (ns)	-5.27	-0.53	5784
NHPT**	0.67	0.58-0.76	0.51 (ns)	5.32	0.53	444
Mental health						
Cognitive function						
FIMcf	0.54	0.44-0.64	-1.41	-6.26	-0.63	80
PASAT3	0.60	0.50-0.69	-0.77 (ns)	26.45	2.77	4816
Emotional well-being						
DIPpsy	0.60	0.50-0.70	-1.11 (ns)	-16.68	-1.68	922
SF36mh	0.55	0.45-0.65	-2.48 (ns)	-28.44	-2.86	537
Social functioning						
DIPsoc	0.64	0.55-0.74	-2.16	-10.27	-1.03	92
RAPocc	0.69	0.61-0.78	-8.40	-26.89	-2.69	42
SF36re	0.53	0.43-0.63	-4.79 (ns)	-74.24	-7.46	980
SF36rp	0.61	0.51-0.71	-12.29	-89.05	-8.95	214
SF36sf	0.60	0.51-0.70	-5.04	-40.71	-4.09	266
General health						
SF36gh	0.51	0.42-0.61	-3.15 (ns)	-30.71	-3.09	388

AUC = area under the ROC curve at three years after baseline with 95%CIs; $MIC_{deterioration}$ = Minimally Important Change; $SRC_{individual}$ = Smallest Real Change at individual level; SRC_{group} = Smallest Real Change at group level, based on 26 stable patients at six months; GRS = Global Rating Scale; ns = not significantly different from 0. * Patients per group, calculation based on a significance level of 0.05 and a power of 0.8. ** MSFC, SaGAS, NHPT, TWT data not transformed into a 0-100 point scale.

from the *patient's* perspective are the SF36pf [0.75 (0.67-0.84)], the DIPmob [0.73 (0.65-0.82)], the RAPocc [0.73 (0.64-0.81)], the DIPself [0.70 (0.62-0.79)] and the EDSS [0.70 (0.62-0.79)]. Of these, only the EDSS does not fulfil the criterion MIC-$P_{deterioration}$ > SRC_{group}. The five most responsive outcome measures to detect *deterioration* (AUC) from the *clinician's* perspective are the SaGAS [0.72 (0.63-0.81)], the SF36pf [0.72 (0.63-0.80)], the MSFC [0.71 (0.62-0.80)], the RAPocc [0.69 (0.61-0.78)] and the TWT [0.69 (0.59-0.78)]. Of these, only the SF36pf and the RAPocc have a MIC-$C_{deterioration}$ > SRC_{group}.

The results for *improvement* are less clear, because of the small percentage of patients in the slightly improved groups (data not shown). The MIC was either very small or did not significantly differ from zero. Therefore, it was not possible to compare the results with the SRC. Consequently, we can only look at the *relative* responsiveness by comparing the AUCs. From the *patient's* perspective, the highest AUCs were found for the EDSS [0.78 (95%CI 0.70-0.87)], the DIP sub-scale mobility [0.73 (95%CI 0.64-0.85)], the FIM sub-scale motor function [0.71 (0.63-0.80)], the SF36pf [0.71 (95%CI 0.62-0.80)], and the RAPocc [0.71 (95%CI 0.62 -0.82)]. From the *clinician's* perspective, the highest AUCs were found for the RAPocc [0.79 (95%CI 0.63-0.95)], the SF36pf [0.77 (95%CI 0.64-0.90)], the FIMmf [0.74 (95%CI 0.62-0.86)], the FIMcf [0.74 (95%CI 0.59-0.90)], and the RAPmob [0.72 (95%CI 0.58-0.87)]. Irrespective of the external criterion that is applied, the most responsive outcome measures to detect *improvement* are the FIMmf, the SF36pf, the RAPocc and the EDSS. However, the criterion MIC > SRC could not be evaluated for any of the measures.

Discussion

In the early stages of MS, the two most useful evaluative outcome measures to detect *deterioration*, and that perform well irrespective of the external criterion that is applied, are the SF36pf for the physical functioning domain (mobility), and the RAPocc for the social functioning domain. Both measures have a MIC > SRC_{group}, which

makes them suitable for application in clinical research. However, none of the outcome measures that we studied had a MIC > $SRC_{individual}$, which means that the reliability demands that warrant application at individual patient level are not met.

The selection of an outcome measure is not only guided by its responsiveness. It is also important to select an outcome measure that really measures the phenomena of interest. Therefore, we categorized the outcome measures that we have studied into five domains and five sub-domains, which should guide their selection. Before the final selection of an outcome measure, one should study the content of an outcome measure to make sure it measures the variable one is interested in. The measures that perform best in the other domains are the DIPpsy (mental health domain, emotional well-being) and the SF36gh (general health domain), but none of the disease-specific outcome measures fulfilled our selection criteria.

We were looking for an outcome measure with a performance that did not depend on the required perspective. Finding such an outcome measure would increase our confidence in this measure, because it would imply that the results obtained with this measure have the same meaning for both the clinician and the patient. However, it might be very legitimate to emphasize one or both perspectives depending on the research aim. For more basic research purposes reliance on examiner-driven outcomes might be fully acceptable. But for more clinically oriented research questions, *i.e.* studies that are interested in the effects on patients, such as clinical trials, reliance on examiner-driven assessments only is not sufficient. In these studies one should also include patient-driven outcome measures, because that is the only way to show benefit for patients. For the evaluation of this kind of clinically oriented research it would be very valuable to have a (primary) outcome measure available which evaluative ability is independent of the chosen perspective (patient vs. examiner), because only then the MIC is the same for the patient and the examiner, which facilitates the interpretation of this research.

An important strength of this study is the simultaneous evaluation of several outcome measures that are frequently used in MS research. Scores were collected for 23 (sub-scales of) outcome measures in the same patients and in the same way. This enables a direct comparison of the outcome measures, and facilitates interpretation of the results. Information about the responsiveness of outcome measures is often derived from several studies with different designs, different populations, different anchors, and different outcome measures. This hampers the selection of the most responsive outcome measure, because no direct comparison can be made.

The *relative* responsiveness is quite independent of the particular approach to the evaluation of responsiveness.[20] We chose the approach presented in this article for two reasons. First of all, we aimed to identify the most responsive outcome measures by comparing the outcome measures on the basis of the AUC (*relative* responsiveness). Second, we tried to obtain data that would facilitate the interpretation of score changes in future MS studies. The interpretation depends on two aspects of the score change: (i) what is a minimally important change, and (ii) is the instrument capable of measuring this change? We have used the MIC as a measure of minimally important change, and the SRC to estimate the ability of a measure to detect this change. From our results we conclude that our strategy worked well for the analysis of changes in the direction of deterioration, because we were able to clearly show the *relative* responsiveness, and provide clear data that facilitate the interpretation of score changes. However, the results with regard to changes in the direction of improvement are inconclusive, due to the small number of patients in this category.

Another aspect of this study that deserves some attention is the analysis of repeated measures. We made optimal use of the longitudinal data by applying longitudinal data-analysis techniques, which reduces the standard error of our estimates. Moreover, we constructed a regression model that enabled us to estimate the MIC for deterioration and improvement in one model.

The possibility of this study to show improvement is limited by its design, because recruiting recently diagnosed patients, who are only mildly disabled, implies a limitation in the possibility to improve. Therefore, our results for improvement are not as clear as those for deterioration. However, despite this limitation, the study does provide some preliminary evidence that the $MIC_{deterioration}$ and the $MIC_{improvement}$ are not necessarily equal.[33]

A well-known problem in studies of anchor-based responsiveness is the choice of the external criterion to define change.[32] Norman *et al.* [73] compared two methods to assess responsiveness with each other: (i) an effective therapy as construct for change, and (ii) a retrospective method to assess change using an GRS. In this direct comparison the GRS performs worse than the effective therapy as external criterion. The problem with the generalization of these results is that there is often not an effective therapy available. Particularly in longitudinal cohort studies, such as ours, we cannot rely on an effective therapy. There are ways to use effective therapy as construct for change in MS by applying outcome measures in patients that were treated for a relapse with corticosteroids. A major problem in these studies is that one is looking at improvements. It is absolutely not certain that these result can subsequently be used in studies that look at deterioration.

Because a gold standard for change is lacking, we had to rely on other methods to define change. Roughly speaking, there are three constructs for the evaluation of change in MS: data obtained from repeated Magnetic Resonance Imaging (MRI) studies, the EDSS as the most frequently used clinical outcome measure, and a GRS which emphasizes the perspective of the patient. Our main focus in this study was on disability and quality of life. Therefore, using MRI data as a construct for change is not appealing, since it only offers information at the level of pathological changes, which are only remotely related to disability and even less related to quality of life. The EDSS has limitations with regard to its validity and reliability, which might make it relatively unsuitable as an

external criterion for change. However, despite this criticism, it is a scale that is very well known among clinicians. It is, in fact, so well known that a description of a study population is not complete without EDSS data. Therefore, we used the EDSS to determine important change from a clinician's point of view. Because the first question of a clinician during a visit often is a global rating: 'How are you doing since the last visit', and because a stronger external criterion is lacking, we used a GRS to emphasize the perspective of the patient.

A global rating requires that patients are able to mentally subtract a previous situation from the present situation.[61,74] Criticism about the use of a global rating scale concerns the fact that this rating has often been found to show stronger associations with the present situation than with the previous situation.[58] In an attempt to overcome this problem, we coupled the previous situation to an important life-event for the patient. In this way, we tried to facilitate the mental subtraction, and hoped for more equal associations of the GRS with the previous and the present situation. We considered the time of diagnosis as an important life-event. Because in our study patients were not diagnosed until some time after their exacerbation and because the mean time between diagnosis and first measurement is relatively short (3.5 months), we decided that is was valid to use it as reference point. Our strategy was partly successful. The mean correlation coefficient between the GRS at three years and the outcome measures at baseline was 0.26 (range 0.15-0.43), at six months it was 0.30 (range 0.14-0.44), at one year it was 0.33 (range 0.14-0.49), at two years it was 0.37 (range 0.09-0.56), and at three years it was 0.40 (range 0.14-0.59).

Another point of discussion about the use of the GRS as external criterion is the choice of the cut-off point used for the calculation of the MIC. We decided to use the category 'slightly deteriorated' or 'slightly improved' as indicator of minimally important change. In our opinion, the next category ('much deteriorated' or 'much improved') is, at least semantically, not equivalent to mini-

mally important change. Others have argued that using 'much deteriorated' or 'much improved' is more appropriate than 'slightly deteriorated' or 'slightly improved', because the latter two categories are often used by patients who are reluctant to classify themselves as stable, while their situation would justify this classification.[75] We performed a sensitivity analysis (data not shown), with the category 'much deteriorated' as cut-off, and compared the MIC-P and the MIC-P estimates obtained in this sensitivity analysis (MIC-P_{sens}) with the MIC-C. For 17 outcome measures the MIC-P was closer to the MIC-C than the MIC-P_{sens}, indicating that there is a greater correspondence between the MIC-P and the MIC-C than between the MIC-P_{sens} and the MIC-C, which supports the use of the category 'slightly deteriorated' as cut-off in this sample. In future studies it might be useful to add extra categories to the GRS between 'slightly' and 'much', for example by using 'deteriorated' and 'improved' on their own, and to use these categories to determine the MIC. This might lessen the (semantical) gap between 'slightly' and 'much', and might aid patients who are reluctant to use the category 'stable', without influencing the estimation of the MIC.

Recently, Solari *et al.*[76] studied the practice effects of the MSFC and suggested that, to improve efficiency, one prebaseline administration of TWT, three of PASAT and four of NHPT are needed. Their study consisted of repeated administrations of the tests in one day. What their results mean for repeated MSFC measurements with intervals of six months or longer, such as our study, is not immediately clear. Will you never loose your ability to perform the PASAT or NHPT once you have mastered it, or do you again need some prebaseline administrations after you have not been performing the PASAT or NHPT for some time? For the components of the MSFC and the SaGAS we used the same test protocol at each measurement. The NHPT and the TWT were conducted twice. For the TWT this is sufficient, for the NHPT two additional administrations would have been better. The PASAT was always administered

once, but in any case after at least one practice trial, as described in the MSFC manual. Although the interval between subsequent measurements was at least six months, we cannot rule out a practice effect. Ignoring a possibly present practice effect will lead to inflated measures of responsiveness in the direction of deterioration for the NHPT and PASAT, because the measured change in cognitive or upper limb function is smaller than the real change. The opposite would occur for the measures of responsiveness in the direction of improvement, because the measured improvement in cognitive function is larger than the real improvement.

Although we were able to identify the most responsive outcome measures and to show, for several of these outcome measures, that the signal (MIC) exceeds the noise (SRC_{group}), it should be noted that our results are not automatically applicable to all patients with MS. In general, our population was only mildly disabled, and had a disease duration of just over three years at the end of the study. The results of this study can therefore be used in early intervention studies. With the positive effects of disease modifying treatments, patients will be mildly disabled for a longer period. Future trials will have to compare newly developed treatments with the current disease modifying treatments. Showing differences in effectiveness in these studies will increasingly suffer from power problems. In comparative studies an outcome measure should be able to show *differences* between *longitudinal changes* of two (or more) groups (arms of a trial), which is probably more difficult than showing changes within one group only. In our opinion this a requirement that can only be fulfilled when an outcome measure is already capable of detecting longitudinal changes. Our results clearly show that some of the outcome measures that we have studied, and that are not regularly used in trials, are more suitable to evaluate changes than others. Future responsiveness studies should focus on more severely disabled populations and populations with a longer duration of the disease.

None of the outcome measures used in this study could detect

important change in individual patients. Outcome measures that might be useful should have a relatively low $SRC_{individual}$. This point has already been acknowledged in relation to the MSFC. Several authors have stated that a change of 20% for the components of the MSFC is required to exceed measurement error[77,78] and that changes for the MSFC and SaGAS should be >0.5.[8,79] Depending on the external criterion used, we found that in our sample a change of 2.6-3.0 s (40% of baseline) for the TWT and 2.8-5.3 s (13% of baseline) for the NHPT is required to exceed measurement error. In our sample, changes in MSFC and SaGAS should exceed 0.54-0.72 and 0.25-0.44, respectively, in order to indicate significant change. However, MSFC scores should be interpreted with caution, because it is not evident from the total score which component contributes most to the total score. The differences between results reported in the literature[8,77-79] and our results might be explained by our study design. We recruited recently diagnosed patients, whereas in the other studies the patients had the disease for a various length of time. Furthermore, we used a fixed interval of six months between visits to identify the stable patients, whereas the other studies used a five-day or a variable interval. The design of the present study matches usual patient care, which increases the validity of our results, but, unfortunately, leads to the conclusion that the outcome measures in this study are not suitable for detecting change within a few years in individual, recently diagnosed, patients.

References

1. Noseworthy JH. Clinical scoring methods for multiple sclerosis. Ann Neurol 1994; 36:S80-S85.
2. Sharrack B, Hughes RA. The Guy's Neurological Disability Scale (GNDS): a new disability measure for multiple sclerosis. Mult Scler 1999; 5:223-233.
3. Hobart J, Freeman J, Thompson A. Kurtzke scales revisited: the application of psychometric methods to clinical intuition. Brain 2000; 123:1027-1040.
4. Cutter GR, Baier ML, Rudick RA, Cookfair DL, Fischer JS, Petkau J *et al.* Development of a multiple sclerosis functional composite as a clinical trial outcome measure. Brain 1999; 122:871-882.

5. Fischer JS, Rudick RA, Cutter GR, Reingold SC. The Multiple Sclerosis Functional Composite Measure (MSFC): an integrated approach to MS clinical outcome assessment. National MS Society Clinical Outcomes Assessment Task Force. Mult Scler 1999; 5:244-250.
6. Miller DM, Rudick RA, Cutter GR, Baier M, Fischer JS. Clinical significance of the multiple sclerosis functional composite: relationship to patient-reported quality of life. Arch Neurol 2000; 57:1319-1324.
7. Cohen JA, Cutter GR, Fischer JS, Goodman AD, Heidenreich FR, Jak AJ *et al.* Use of the multiple sclerosis functional composite as an outcome measure in a phase 3 clinical trial. Arch Neurol 2001; 58:961-967.
8. Vaney C, Vaney S, Wade DT. SaGAS, the Short and Graphic Ability Score: an alternative scoring method for the motor components of the Multiple Sclerosis Functional Composite. Mult Scler 2004; 10:231-242.
9. Hobart J, Lamping D, Fitzpatrick R, Riazi A, Thompson A. The Multiple Sclerosis Impact Scale (MSIS-29): a new patient-based outcome measure. Brain 2001; 124:962-973.
10. Granger CV, Cotter AC, Hamilton BB, Fiedler RC, Hens MM. Functional assessment scales: A study of persons with multiple sclerosis. Arch Phys Med Rehabil 1990; 71:870-875.
11. Freeman JA, Hobart JC, Langdon DW, Thompson AJ. Clinical appropriateness: a key factor in outcome measure selection: the 36 item short form health survey in multiple sclerosis. J Neurol Neurosurg Psychiatry 2000; 68:150-156.
12. Kidd D, Howard RS, Losseff NA, Thompson AJ. The benefit of inpatient neurorehabilitation in multiple sclerosis. Clin Rehabil 1995; 9:198-203.
13. Ottenbacher KJ, Hsu Y, Granger CV, Fiedler RC. The reliability of the functional independence measure: a quantitative review. Arch Phys Med Rehabil 1996; 77:1226-1232.
14. Hobart JC, Lamping DL, Freeman JA, Langdon DW, McLellan DL, Greenwood RJ *et al.* Evidence-based measurement: which disability scale for neurologic rehabilitation? Neurology 2001; 57:639-644.
15. Jonsson A, Dock J, Ravnborg MH. Quality of life as a measure of rehabilitation outcome in patients with multiple sclerosis. Acta Neurologica Scandinavica 1996; 93:229-235.
16. Lankhorst GJ, Jelles F, Smits RCF, Polman CH, Kuik DJ, Pfennings LEMA *et al.* Quality of life in multiple sclerosis: The disability and impact profile (DIP). J Neurol 1996; 243:469-474.
17. Pfennings LE, van der Ploeg HM, Cohen L, Bramsen I, Polman CH, Lankhorst GJ *et al.* A health-related quality of life questionnaire for multiple sclerosis patients. Acta Neurol Scand 1999; 100:148-155.

18. Cohen L, Pouwer F, Pfennings LE, Lankhorst GJ, van der Ploeg HM, Polman CH *et al.* Factor structure of the Disability and Impact Profile in patients with multiple sclerosis. Qual Life Res 1999; 8:141-150.
19. van der Putten JJ, Hobart JC, Freeman JA, Thompson AJ. Measuring change in disability after inpatient rehabilitation: comparison of the responsiveness of the Barthel index and the Functional Independence Measure. J Neurol Neurosurg Psychiatry 1999; 66:480-484.
20. Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Qual Life Res 2003; 12:349-362.
21. de Vet HCW, Bouter LM, Bezemer PD, Beurskens AJ. Reproducibility and responsiveness of evaluative outcome measures. Theoretical considerations illustrated by an empirical example. Int J Technol Assess Health Care 2001; 17:479-487.
22. Koziol JA, Lucero A, Sipe JC, Romine JS, Beutler E. Responsiveness of the Scripps neurologic rating scale during a multiple sclerosis clinical trial. Can J Neurol Sci 1999; 26:283-289.
23. Schwid SR, Goodman AD, Apatoff BR, Coyle PK, Jacobs LD, Krupp LB *et al.* Are quantitative functional measures more sensitive to worsening MS than traditional measures? Neurology 2000; 55:1901-1903.
24. Hoogervorst EL, Kalkers NF, van Winsen LML, Uitdehaag BMJ, Polman CH. Differential treatment effect on measures of neurologic exam, functional impairment and patient self-report in multiple sclerosis. Mult Scler 2001; 7:335-339.
25. Patzold T, Schwengelbeck M, Ossege LM, Malin JP, Sindern E. Changes of the MS functional composite and EDSS during and after treatment of relapses with methylprednisolone in patients with multiple sclerosis. Acta Neurol Scand 2002; 105:164-168.
26. Riazi A, Hobart JC, Lamping DL, Fitzpatrick R, Thompson AJ. Evidence-based measurement in multiple sclerosis: the psychometric properties of the physical and psychological dimensions of three quality of life rating scales. Mult Scler 2003; 9:411-419.
27. McGuigan C, Hutchinson M. The multiple sclerosis impact scale (MSIS-29) is a reliable and sensitive measure. J Neurol Neurosurg Psychiatry 2004; 75:266-269.
28. Uitdehaag BMJ, Ader HJ, Roosma TJ, de Groot V, Kalkers NF, Polman CH. Multiple sclerosis functional composite: impact of reference population and interpretation of changes. Mult Scler 2002; 8:366-371.
29. Hobart JC, Riazi A, Lamping DL, Fitzpatrick R, Thompson AJ. Improving the

evaluation of therapeutic interventions in multiple sclerosis: development of a patient-based measure of outcome. Health Technol Assess 2004; 8:1-60.

30. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003; 56:395-407.
31. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol 2000; 53:459-468.
32. Cella D, Eton DT, Lai JS, Peterman AH, Merkel DE. Combining anchor and distribution-based methods to derive minimal clinically important differences on the Functional Assessment of Cancer Therapy (FACT) anemia and fatigue scales. J Pain Symptom Manage 2002; 24:547-561.
33. Cella D, Hahn EA, Dineen K. Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening. Qual Life Res 2002; 11:207-221.
34. Lydick E, Epstein RS. Interpretation of quality of life changes. Qual Life Res 1993; 2:221-226.
35. Schmitt JS, Di Fabio RP. Reliable change and minimum important difference (MID) proportions facilitated group responsiveness comparisons using individual threshold criteria. J Clin Epidemiol 2004; 57:1008-1018.
36. Schunemann HJ, Griffith L, Jaeschke R, Goldstein R, Stubbing D, Guyatt GH. Evaluation of the minimal important difference for the feeling thermometer and the St. George's Respiratory Questionnaire in patients with chronic airflow obstruction. J Clin Epidemiol 2003; 56:1170-1176.
37. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis 1987; 40:171-178.
38. Poser CM, Paty DW, Scheinberg L, McDonald WI, Davis FA, Ebers GC *et al.* New diagnostic criteria for multiple sclerosis: guidelines for research protocols. Ann Neurol 1983; 13:227-231.
39. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology 1983; 33:1444-1452.
40. Whitaker JN, Mcfarland HF, Rudge P, Reingold SC. Outcomes assessment in multiple sclerosis clinical trials: a critical analysis. Mult Scler 1995; 1:37-47.
41. Rudick R, Antel J, Confavreux C, Cutter G, Ellison G, Fischer J *et al.* Clinical outcomes assessment in multiple sclerosis. Ann Neurol 1996; 40:469-479.
42. Cohen JA, Fischer JS, Bolibrush DM, Jak AJ, Kniker JE, Mertz LA *et al.* Intrarater and interrater reliability of the MS functional composite outcome measure. Neurology 2000; 54:802-806.
43. Kalkers NF, de Groot V, Lazeron RH, Killestein J, Ader HJ, Barkhof F *et al.* MS functional composite: relation to disease phenotype and disability strata. Neurology 2000; 54:1233-1239.

44. Kalkers NF, Bergers E, Castelijns JA, van Walderveen MA, Bot JC, Ader HJ *et al.* Optimizing the association between disability and biological markers in MS. Neurology 2001; 57:1253-1258.
45. Hoogervorst EL, van Winsen LM, Eikelenboom MJ, Kalkers NF, Uitdehaag BM, Polman CH. Comparisons of patient self-report, neurologic examination, and functional impairment in MS. Neurology 2001; 56:934-937.
46. Lyle RC. A performance test for assessment of upper limb function in physical rehabilitation treatment and research. Int J Rehabil Res 1981; 4:483-492.
47. van der Lee JH, de Groot V, Beckerman H, Wagenaar RC, Lankhorst GJ, Bouter LM. The intra- and interrater reliability of the action research arm test: a practical test of upper extremity function in patients with stroke. Arch Phys Med Rehabil 2001; 82:14-19.
48. Laman H, Lankhorst GJ. Subjective weighting of disability: An approach to quality of life assessment in rehabilitation. Disabil Rehabil 1994; 16:198-204.
49. Marolf MV, Vaney C, Konig N, Schenk T, Prosiegel M. Evaluation of disability in multiple sclerosis patients: A comparative study of the Functional Independence Measure, the extended Barthel Index and the Expanded Disability Status Scale. Clin Rehabil 1996; 10:309-313.
50. van Bennekom CAM, Jelles F, Lankhorst GJ, Bouter LM. The Rehabilitation Activities Profile: A validation study of its use as a disability index with stroke patients. Arch Phys Med Rehabil 1995; 76:501-507.
51. van Bennekom CAM, Jelles F, Lankhorst GJ, Bouter LM. Responsiveness of the Rehabilitation Activities Profile and the Barthel Index. J Clin Epidemiol 1996; 49:39-44.
52. Collen FM, Wade DT, Robb GF, Bradshaw CM. The Rivermead Mobility Index: a further development of the Rivermead Motor Assessment. Int Disabil Stud 1991; 13:50-54.
53. Forlander DA, Bohannon RW. Rivermead Mobility Index: a brief review of research to date. Clin Rehabil 1999; 13:97-100.
54. Hsieh CL, Hsueh IP, Mao HF. Validity and responsiveness of the Rivermead Mobility Index in stroke patients. Scand J Rehabil Med 2000; 32:140-142.
55. Antonucci G, Aprile T, Paolucci S. Rasch analysis of the Rivermead Mobility Index: a study using mobility measures of first-stroke inpatients. Arch Phys Med Rehabil 2002; 83:1442-1449.
56. Brunet DG, Hopman WM, Singer MA, Edgar CM, MacKenzie TA. Measurement of health-related quality of life in multiple sclerosis patients. Can J Neurol Sci 1996; 23:99-103.
57. Vickrey BG, Hays RD, Harooni R, Myers LW, Ellison GW. A health-related quality of life measure for multiple sclerosis. Qual Life Res 1995; 4:187-206.

58. Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition ratings. J Clin Epidemiol 2002; 55:900-908.
59. Bessette L, Sangha O, Kuntz KM, Keller RB, Lew RA, Fossel AH *et al.* Comparative responsiveness of generic versus disease-specific and weighted versus unweighted health status measures in carpal tunnel syndrome. Med Care 1998; 36:491-502.
60. Stucki G, Liang MH, Fossel AH, Katz JN. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. J Clin Epidemiol 1995; 48:1369-1378.
61. Liang MH. Evaluating measurement responsiveness. J Rheumatol 1995; 22:1191-1192.
62. Juniper EF, Guyatt GH, Willan A, Griffith LE. Determining a minimal important change in a disease-specific Quality of Life Questionnaire. J Clin Epidemiol 1994; 47:81-87.
63. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989; 10:407-415.
64. Beurskens AJHM, de Vet HCW, Köke AJA. Responsiveness of functional status in low back pain: a comparison of different instruments. Pain 1996; 65:71-76.
65. van der Windt DA, van der Heijden GJ, de Winter AF, Koes BW, Devillé W, Bouter LM. The responsiveness of the Shoulder Disability Questionnaire. Ann Rheum Dis 1998; 57:82-87.
66. Mancuso CA, Peterson MG. Different methods to assess quality of life from multiple follow-ups in a longitudinal asthma study. J Clin Epidemiol 2004; 57:45-54.
67. Wyrwich KW, Tierney WM, Wolinsky FD. Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. J Clin Epidemiol 1999; 52:861-873.
68. Noseworthy JH, Vandervoort MK, Wong CJ, Ebers GC. Interrater variability with the Expanded Disability Status Scale (EDSS) and Functional Systems (FS) in a multiple sclerosis clinical trial. The Canadian Cooperation MS Study Group. Neurology 1990; 40:971-975.
69. Goodkin DE, Cookfair D, Wende K, Bourdette D, Pullicino, Scherokman B *et al.* Inter- and intrarater scoring agreement using grades 1.0 to 3.5 of the Kurtzke Expanded Disability Status Scale (EDSS). Multiple Sclerosis Collaborative Research Group. Neurology 1992; 42:859-863.
70. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 1986; 42:121-130.
71. Pfennings LE, van der Ploeg HM, Cohen L, Polman CH. A comparison of responsiveness indices in multiple sclerosis patients. Qual Life Res 1999; 8:481-489.

72. Beckerman H, Roebroeck ME, Lankhorst GJ, Becher JG, Bezemer PD, Verbeek ALM. Smallest real difference, a link between reproducibility and responsiveness. Qual Life Res 2001; 10:571-578.
73. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. Journal of Clinical Epidemiology 1997; 50:869-879.
74. Stratford PW, Binkley FM, Riddle DL. Health status measures: strategies and analytic methods for assessing change scores. Phys Ther 1996; 76:1109-1123.
75. Ostelo RW, de Vet HC. Clinically important outcomes in low back pain. Best Pract Res Clin Rheumatol 2005; 19:593-607.
76. Solari A, Radice D, Manneschi L, Motti L, Montanari E. The multiple sclerosis functional composite: different practice effects in the three test components. J Neurol Sci 2005; 228:71-74.
77. Schwid SR, Goodman AD, McDermott MP, Bever CF, Cook SD. Quantitative functional measures in MS: what is a reliable change? Neurology 2002; 58:1294-1296.
78. Kaufman M, Moyer D, Norton J. The significant change for the Timed 25-foot Walk in the multiple sclerosis functional composite. Mult Scler 2000; 6:286-290.
79. Hoogervorst EL, Zwemmer JN, Jelles B, Polman CH, Uitdehaag BMJ. Multiple Sclerosis Impact Scale (MSIS-29): relation to established measures of impairment and disability. Mult Scler 2004; 10:569-574.

4

The initial course of daily functioning in multiple sclerosis: a three-year follow-up study

Vincent de Groot, Heleen Beckerman, Gustaaf J. Lankhorst, Chris H. Polman & Lex M. Bouter

Mult Scler 2005; 11:713-718

Abstract

Objectives. To study the initial course of daily functioning in multiple sclerosis (MS).

Methods. A cohort of 156 recently diagnosed patients was prospectively followed for three years (five measurements). Domains of interest were neurological deficits, physical functioning, mental health, social functioning and general health. An a priori distinction was made between a relapse onset group (n=128) and a non-relapse onset group (n=28).

Results. At baseline, neurological deficits are relatively minor for most patients, 26.3% have aberrant physical functioning scores, 38.5% have aberrant social functioning scores, 9.0% have aberrant mental health scores and 25.0% have aberrant general health scores. The neurological deficits and physical functioning deteriorated significantly over time. This deterioration was more pronounced and clinically relevant in the non-relapse onset group only. Mental health showed a significant, but not clinically relevant deterioration over time. Social functioning and general health showed non-significant effects for time.

Conclusions. In the initial stage of MS, when neurological deficits are relatively minor and mental health is relatively unaffected, patients in both groups experience limitations in daily functioning. Patients in the non-relapse onset group have progressive neurological symptoms that are accompanied by progressive limitations in physical functioning, but not by progressive limitations in the other domains.

Introduction

Most longitudinal studies on the clinical course of multiple sclerosis (MS) in large study populations have only used the Expanded Disability Status Scale (EDSS) as outcome measure,[1-7] which is subject to criticism.[8-10] A clear limitation is that it combines impairments and disability into one scale. Moreover, the instrument is heavily biased towards locomotor function, and does not cover other relevant domains of functioning. Studies that address other domains of functioning are predominately cross-sectional,[11] thereby not providing insight into the course of MS in these domains. Although the information obtained in these studies is certainly of value, longitudinal studies of carefully and comprehensively documented cohorts of patients with MS would improve our knowledge on the course of MS in relevant domains of daily functioning.

We studied the initial course of MS in the domains of neurological deficits, physical functioning, mental health, social functioning and general health for the relapse onset group [RO: Relapsing Remitting MS (RRMS) and Secondary Progressive MS (SPMS)] and the non-relapse onset group [NRO: Primary Progressive MS (PPMS) and patients for whom the type is unknown at six months] in the first three years.

Methods

All consecutive potentially eligible patients visiting the outpatient clinics of five participating neurology departments were invited to participate in the study. A cohort of 156 MS patients, diagnosed less than six months previously and aged 16-55 years, was recruited and prospectively followed for three years. Diagnoses were made according to the Poser-criteria for definite MS.[12] Patients with other neurological disorders, systemic or malignant neoplastic diseases were excluded. Measurements took place at baseline, after six months, and after one, two and three years. In case of a relapse, the measurements were postponed for a few weeks until the relapse had subsided. Patients were visited at home in order to minimise drop-out.

The cohort was sub-divided into two a priori defined groups on the basis of disease type, determined six months after inclusion in the study.[13] The RO group consisted of patients with RRMS or SPMS, and was the reference category in the analysis. The NRO group consisted of patients with PPMS and patients with MS for which the disease type was unknown at six months.

The EDSS,[14] the Functional Independence Measure (FIM)[15] and the Medical Outcome Study Short Form 36 (SF36),[16,17] or their sub-scales, were used to assess the domains of neurological deficits, physical functioning, mental health, social functioning and general health (TABLE 4.1). The FIM and SF36 scores were transformed to a scale that ranged from 0 (worst) to 100 (best). The EDSS was used in its original format, where 0 indicates no neurological deficits and 10 indicates death due to MS.

The EDSS[14] consists of a thorough neurological assessment of the seven neurological systems (visual/optic, brainstem, pyramidal, cerebellar, bowel/bladder, mental and other) and provides information about walking ability, use of walking aids and ability to perform self-care activities. Lower scores on the EDSS are determined with a scoring paradigm based on the scores obtained from the neurological systems, intermediate scores are predominantly based on walking ability, and higher scores are mainly based on the inability to perform self-care activities. Reliability has been shown to be moderate.[10] Therefore, only experienced raters were involved in the scoring.

The FIM consists of a motor function (FIMmf, 13 items) and a cognitive function (FIMcf, 5 items) sub-scale.[15] The items address the activities of daily living, and are scored on the basis of a semi-structured interview. The validity of the FIM has been established for use in inpatient and outpatient rehabilitation settings,[18,19] and its reliability is good.[18,19]

The SF36[16,17] is a questionnaire that assesses eight domains [physical functioning (SF36pf), mental health (SF36mh), bodily pain, vitality, social functioning (SF36sf), role physical (SF36rp), role

TABLE 4.1 Domains of interest, (sub-scales of) outcome measures used and percentage (95%CI) of MS patients with aberrant scores at baseline.

Domain	Outcome measures	Abbreviation	Percentage with aberrant scores at baseline
Neurological deficits	Expanded Disability Status Scale	EDSS	96.8 (95.4-98.2)%*
Physical functioning	FIM motor function	FIMmf	74.4 (70.9-87.9)%**
	SF36 physical functioning	SF36pf	26.3 (19.4-33.2)%***
Mental health	FIM cognitive function	FIMcf	64.7 (60.9-68.5)%**
	SF36 mental health	SF36mh	9.0 (4.5-13.5)%***
Social functioning	SF36 role physical	SF36rp	38.5 (30.8-46.1)%***
	SF36 role emotional	SF36re	12.2 (7-17.3)%***
	SF36 social functioning	SF36sf	12.8 (7.6-18.1)%***
General health	SF36 general health perception	SF36gh	25.0 (18.2-31.8)%***

FIM = Functional Independence Measure; SF36 = Medical Outcome Study Short Form 36; *Percentage scoring ≥ 1; ** Percentage scoring < 100; *** Percentage scoring worse than 1.96 standard deviations below the mean of the reference data obtained from Aaronson et al.[17]

emotional (SF36re), general health perception (SF36gh)]. Its validity and reliability have been extensively studied.[16,17] For MS it is not recommended to calculate physical and mental component scores, because scaling assumptions would be violated.[20]

To study the differences in SF36 scores between a healthy population and the study population, we used reference data on an age-matched healthy Dutch reference population, derived from Aaronson *et al.*[17] The cut-off was set at 1.96 standard deviations below the mean of the reference population. EDSS scores >0 and FIM scores <100 were considered aberrant. We calculated 95% confidence intervals (95%CI) around the proportion of patients with aberrant scores.

For the longitudinal analysis of the domains of neurological deficits, physical functioning, mental health and general health perception, data is presented on six models, using raw data for the graphs and *linear* generalised estimating equations (GEE)[21] from

the Statistical Package for Interactive Data Analysis (SPIDA) version 6.05 from the Statistical Computing Laboratory for the analysis. The correlation structure was chosen on the basis of the correlation matrix of the outcome measures, and set at exchangeable (*i.e.* correlation coefficients between the first and successive measurements are approximately equal) for all outcomes except the cognitive sub-scale of the FIM that was set at four-dependence (*i.e.* correlation coefficients between the first and successive measurements are progressively smaller). For the domain of social functioning (SF36rp, SF36re and SF36sf) data is presented on three models using raw data for the graphs and *binomial* GEE for the analysis. Because the data showed strong floor and ceiling effects, we distinguished a group scoring within the norm and a group with scores deviating from the norm, using a cut-off of 1.96 standard deviations below the mean of an age-matched Dutch reference population.[17] Time was modelled as a continuous variable, expressed in years. To test for differences in the course of both groups we used an interaction term time x group. The significance level for time, group and time x group was set at 0.05. Determining the minimally clinically important difference (MCID) depends on numerous factors and assumptions.[22] For the present study the MCID was set at a 10% difference for all outcome measures.

Results

TABLE 4.2 shows the baseline characteristics of the patients. Most characteristics comply with the expected pattern: 64% females, approximately 80% with a relapse onset, more females than males in the RRMS group, more males than females in the PPMS group, and more severe neurological deficits in the groups with a progressive disease course.[1,4,6] Seven patients were lost to follow-up (three after one year, one after two years and three after three years) and only 1.9% of the measurements were missing.

TABLE 4.1 shows the proportion of patients with aberrant scores for all outcome measures. Only 3% of the patients have no

TABLE 4.2 Baseline characteristics of patients with multiple sclerosis.

	Relapse onset		Non-relapse onset		
	RR	SP	PP	Not yet known	Total
n (%)	120 (77%)	8 (5%)	25 (16%)	3 (2%)	156 (100%)
Age (SD)	35.5 (8.9)	48.2 (6.7)	43.2 (8.9)	45.5 (6.9)	37.6 (9.5)
Female (%)	84 (70.0%)	3 (37.5%)	11 (44%)	3 (100%)	101 (64%)
Years since diagnosis	0.26 (0.15-0.41)	0.33 (0.24-0.48)	0.28 (0.15-0.33)	0.14 (0.14-0.17)	0.26 (0.15-0.40)
Years since symptoms	1.83 (0.67-4.40)	7.50 (3.35-14.51)	2.10 (1.07-3.15)	3.62 (3.53-4.63)	2.15 (0.79-4.36)
Number of exacerbations	2 (1-3)	2 (1-7)	0 (0-0)	0 (0-0)	2 (1-3)
EDSS	2.0 (2.0-3.0)	3.0 (2.5-3.9)	3.0 (2.5-4.0)	2.5 (2.0-4.0)	2.5 (2.0-3.0)

n (percentage), mean (SD) or median (IQR). RR = relapsing remitting multiple sclerosis; SP = secondary progressive multiple sclerosis; PP = primary progressive multiple sclerosis. EDSS = Expanded Disability Status Scale, original score.

neurological symptoms. Although it seems as if major problems exist on the FIM sub-scales, only 17.3% of patients score less than 90 points on the FIMmf and 10.2% of patients score less than 90 points on the FIMcf. This indicates that current disabilities are relatively minor for most patients. Deviations from normal are most pronounced for the sub-scales SF36pf, SF36rp and SF36gh.

The course of MS for the two groups can be found in FIGURE 4.1 (raw data) and the corresponding results of the GEE analysis can be found in TABLE 4.3. The neurological deficits (EDSS) and physical functioning (FIMmf and SF36pf) deteriorate in the first three years (time is significant: see TABLE 4.3 and FIGURES 4.1A and 4.1B). For the FIMmf there is a difference between the two groups that does not change over time (group is significant: see TABLE 4.3), but for the EDSS and the SF36pf the deterioration is more pronounced in the NRO group (time x group is significant: see TABLE 4.3). In the NRO group the change in EDSS and SF36pf over the first three years exceeds the MCID (EDSS 1.2 and SF36pf 15 units). In the first three years mental health, as measured with the FIMcf, shows a deterioration (3.6 units, statistically significant but smaller than the MCID)

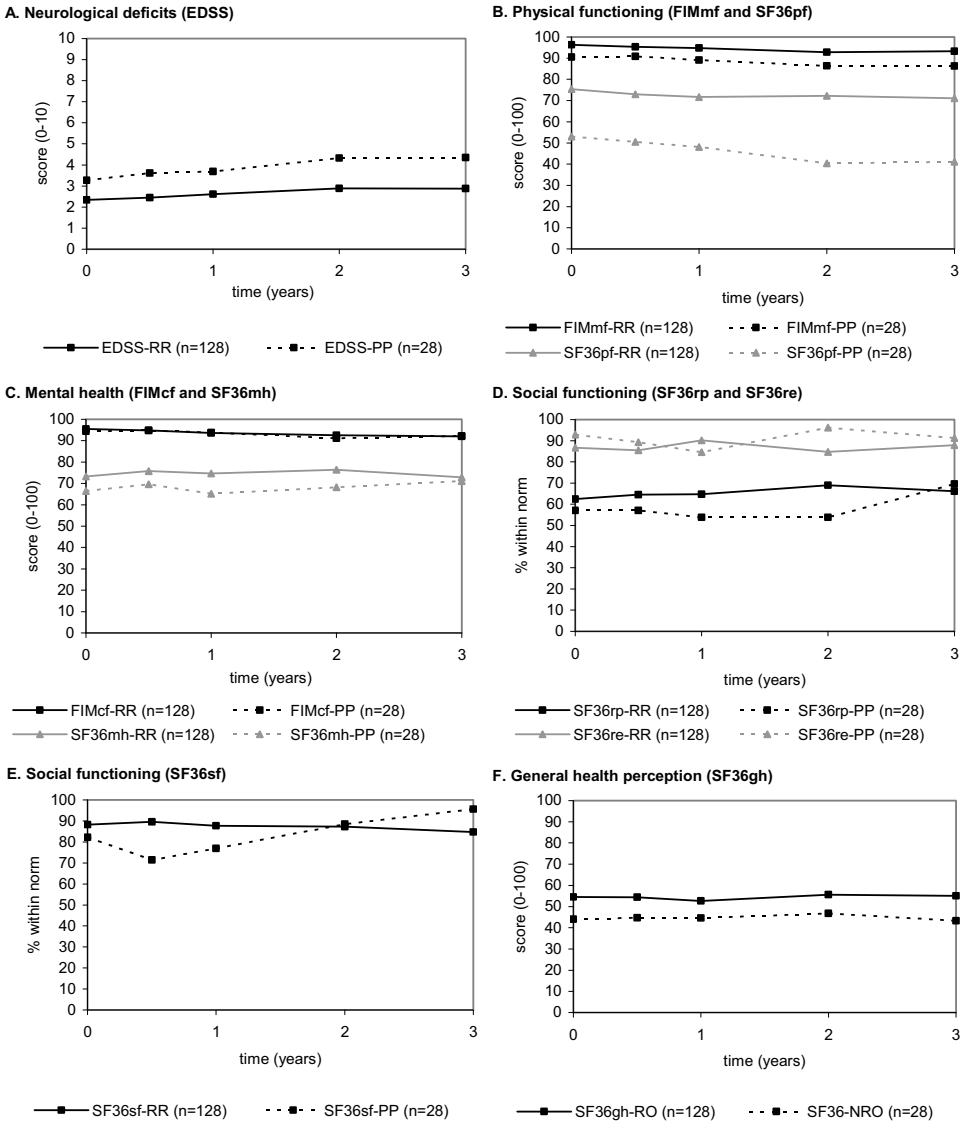


FIGURE 4.1 Initial course of MS in the domains of neurological deficits, physical functioning, mental health, social functioning and general health perception. Graphs are based on raw data. EDSS = Expanded Disability Status Scale; FIMmf = Functional Independence Measure motor function; SF36pf = Medical Outcome Study Short Form 36 items physical functioning; FIMcf = FIM cognitive function; SF36mh = SF36 mental health; SF36rp = SF36 role physical; SF36re = SF36 role emotional; SF36sf = SF36 social functioning; SF36gh = SF36 general health perception; RO = relapse onset group; NRO = non-relapse onset group.

TABLE 4.3 Results (regression coefficient, odds ratios and 95% confidence intervals) of GEE analysis for the different outcome measures.

Outcome measure	Intercept	Time	Group	Time x group
Linear GEE				
EDSS	2.4 (2.2-2.6)	0.2 (0.1-0.3)	1 (0.5-1.4)	0.2 (0.1-0.4)
FIM motor function	96.1 (95.2-96.9)	-1.3 (-1.6- -1.0)	-6.0 (-8.1- -3.8)	
SF36 physical functioning	74.3 (70.5-78.0)	-1.4 (-2.5- -0.2)	-21.9 (-31.4- -12.4)	-3.9 (-6.2- -1.6)
FIM cognitive function	95.0 (94.2-95.8)	-1.0 (-1.4- -0.7)		
SF36 mental health	73.3 (70.7-75.9)	-0.1 (-0.1- -0.1)		
SF36 general health	54.0 (50.9-57.1)	0.4 (-0.5-1.2)	-9.9 (-17.6- -2.2)	
Logistic GEE				
SF36 role physical		1.1 (1.0-1.2)*		
SF36 role emotional		1.0 (0.9-1.2)*		
SF36 social functioning		1.0 (0.8-1.2)*		

GEE = Generalised Estimating Equations; EDSS = Expanded Disability Status Scale; FIM = Functional Independence Measure; SF36 = Medical Outcome Study Short Form 36; Time = time of measurement in years; Group, relapse onset is reference category; Time x Group = interaction term Time with Group. *Odds ratios, logistic GEE models

that is the same in both groups (time is significant: see TABLE 4.3 and FIGURE 4.1C). Mental health, as measured with the SF36mh, does not change significantly. For the other scales no change occurs in the first three years (time is not significant: see TABLE 4.3 and Figures 4.1C, 4.1D1, 4.1D2 and 4.1E).

Scores for a specific group at a specific point in time can be calculated, using the results from TABLE 4.3 in a linear regression formula. As an example, we will calculate the EDSS, FIMmf and SFrp of a patient with MS in the NRO group *two* years after inclusion:

EDSS = 2.4 + 0.2 x time (years) + 1.0 x group (RO = 0, NRO = 1) +
0.2 x time x group = 2.4 + 0.2 x 2 + 1.0 x 1 + 0.2 x 2 x 1 = 4.2

FIMmf = 96.1 – 1.3 x time (in years) – 6.0 x group (RO = 0, NRO = 1)
= 96.1 – 1.3 x 2 – 6.0 x 1 = 87.5

The situation regarding SF36rp is more complex, because the results are presented as an odds ratio (OR). First, the OR from TABLE 4.3 is reverted to the original logistic coefficient by taking the natural log (ln). This logistic coefficient is then multiplied by

time (in years), and finally e is raised to the power of this coefficient:

$$SFrp = e^{\text{time (in years)} \times \ln(1.1)} = e^{\text{time} \times 0.095} = e^{2 \times 0.095} = e^{0.19} = 1.2$$

The number obtained (1.2) is an estimate of the odds (ratio of the probability that the patient deviates from the norm to the probability that he does not) that this patient will have an aberrant social functioning score.

Discussion

At baseline, the domains of physical functioning (SF36pf), social functioning (SF36rp) and general health (SF36gh) are markedly affected. Although both groups are affected, in the domains of physical functioning and general health the NRO group is more severely affected than the RO group, whereas in the domain of social functioning there is no difference between both groups. Surprisingly, mental health is relatively unaffected. These results show that in the initial stage of the disease, when the neurological deficits are relatively minor and mental health is relatively unaffected, patients in both groups do already experience limitations in daily functioning.

In the first three years after diagnosis, the course differs not only between the RO and the NRO group, but also between the five domains. In the domains of neurological deficits and physical functioning the NRO group shows clinically relevant deterioration, whereas the RO groups stays relatively stable. In the domains of mental health, social functioning and general health neither the RO nor the NRO group show any clinically relevant changes. This indicates that patients in the NRO group have progressive neurological symptoms that are accompanied by progressive limitations in physical functioning, but not by progressive limitations in the other domains.

In the later stages of MS, mental health is negatively affected, and there is a relationship between disease severity and mental

health.[23] In contrast to what we expected, we found that mental health was relatively unaffected at baseline or after three years, for which there is no good explanation. Even though the majority of the study population showed only mild neurological symptoms at baseline, we expected that the emotional burden shortly after the diagnosis would have a negative influence on mental health. However, the interval between making the diagnosis and inclusion in the study (maximal six months) may be long enough for patients to recover from an initial deterioration in mental health. Another explanation might be that the outcome measure that was used was not sensitive enough to detect problems in this area.

There are some important strengths of this study. The cohort consists of incident cases of MS, which means that the start of participation in this cohort is clearly defined. Only seven patients were lost to follow-up. Finally, we used a powerful design to study daily functioning. To our knowledge, this is the first longitudinal study that simultaneously assesses several domains of daily functioning. The longitudinal measurements, the concurrent use of several outcome measures at the same points in time, and the use of longitudinal data analysis techniques enable us to make a detailed and comprehensive description of the course of daily functioning in MS.

A potential weakness is the definition of the type of MS. RRMS is relatively easy to recognise, and accounts for the majority of the cases. In practice, PPMS is more difficult to recognise. Furthermore, there is a small sub-group that cannot be classified in the early stages of the disease. During follow-up, it is easier to determine the type of MS, so we choose to dichotomise the patients on the basis of their disease onset type determined six months after inclusion in the study.

Another potential problem in this study is that five patients were classified as SPMS at baseline. This is rather unexpected in this incidence cohort, because this type of MS is normally preceded by RRMS. TABLE 4.2 shows a long time since the first symptoms

for these patients. Looking carefully at their history, it became clear that for all of these patients there was a delay in making the diagnosis, either caused by the patient or the physician. This delay might lead to onset confounding,[24,25] because time since first symptoms is related to disease progression and to conversion of RRMS into SPMS. To study the possibility of onset confounding, we repeated the analysis and adjusted for time since first symptoms at baseline (a logarithmic transformation was applied to obtain a normal distribution). Because none of the coefficients showed a considerable change, it is concluded that onset confounding did not play a major role in the present study.

Although this study clearly shows that in the early phase of MS functioning is already seriously affected, knowledge about the precise mechanisms underlying this limited functioning is scarce. Clinicians might be encouraged to pay special attention to daily functioning in patients who visit their clinic, and explore the possible causes, beside neurological deficits, of the problems in daily functioning. Factors that might contribute to problems in daily functioning might be patient related such as fatigue, personality, depression, and uncertainty about the future, or more related to the environment, such as social support and work related factors. Future studies should focus on the determinants of this limited functioning in order to enhance our understanding of these mechanisms and to provide clinicians with information that can be used in the development of effective treatments.

References

1. Runmarker B, Andersen O. Prognostic factors in a multiple sclerosis incidence cohort with twenty-five years of follow-up. Brain 1993; 116:117-134.
2. Weinshenker BG, Bass B, Rice GPA, Noseworthy J, Carriere W, Baskerville J *et al.* The natural history of multiple sclerosis: A geographically based study. 2. Predictive value of the early clinical course. Brain 1989; 112:1419-1428.
3. Weinshenker BG, Bass B, Rice GPA, Noseworthy J, Carriere W, Baskerville J *et al.* The natural history of multiple sclerosis: A geographically based study. I. Clinical course and disability. Brain 1989; 112:133-146.

4. Weinshenker BG, Rice GPA, Noseworthy JH, Carriere W, Baskerville J, Ebers GC. The natural history of multiple sclerosis: A geographically based study. 3. Multivariate analysis of predictive factors and models of outcome. Brain 1991; 114:1045-1056.
5. Amato MP, Ponziani G. A prospective study on the prognosis of multiple sclerosis. Neurol Sci 2000; 21:S831-S838.
6. Confavreux C, Vukusic S, Adeleine P. Early clinical predictors and progression of irreversible disability in multiple sclerosis: an amnesic process. Brain 2003; 126:770-782.
7. Pittock SJ, Mayr WT, McClelland RL, Jorgensen NW, Weigand SD, Noseworthy JH *et al.* Change in MS-related disability in a population-based cohort: a 10-year follow-up study. Neurology 2004; 62:51-59.
8. Noseworthy JH. Clinical scoring methods for multiple sclerosis. Ann Neurol 1994; 36:S80-S85.
9. Sharrack B, Hughes RA, Soudain S, Dunn G. The psychometric properties of clinical rating scales used in multiple sclerosis. Brain 1999; 122:141-159.
10. Hobart J, Freeman J, Thompson A. Kurtzke scales revisited: the application of psychometric methods to clinical intuition. Brain 2000; 123:1027-1040.
11. Nortvedt MW, Riise T. The use of quality of life measures in multiple sclerosis research. Mult Scler 2003; 9:63-72.
12. Poser CM, Paty DW, Scheinberg L, McDonald WI, Davis FA, Ebers GC *et al.* New diagnostic criteria for multiple sclerosis: guidelines for research protocols. Ann Neurol 1983; 13:227-231.
13. Lublin FD, Reingold SC. Defining the clinical course of multiple sclerosis: results of an international survey. National Multiple Sclerosis Society (USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis. Neurology 1996; 46:907-911.
14. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology 1983; 33:1444-1452.
15. Granger CV, Cotter AC, Hamilton BB, Fiedler RC, Hens MM. Functional assessment scales: A study of persons with multiple sclerosis. Arch Phys Med Rehabil 1990; 71:870-875.
16. Freeman JA, Hobart JC, Langdon DW, Thompson AJ. Clinical appropriateness: a key factor in outcome measure selection: the 36 item short form health survey in multiple sclerosis. J Neurol Neurosurg Psychiatry 2000; 68:150-156.
17. Aaronson NK, Muller M, Cohen PD, Essink-Bot ML, Fekkes M, Sanderman R *et al.* Translation, validation, and norming of the Dutch language version of the SF-36 Health Survey in community and chronic disease populations. J Clin Epidemiol 1998; 51:1055-1068.

18. Kidd D, Stewart G, Baldry J, Johnson J, Rossiter D, Petruckevitch A *et al.* The Functional Independence Measure: a comparative validity and reliability study. Disabil Rehabil 1995; 17:10-14.
19. Ottenbacher KJ, Hsu Y, Granger CV, Fiedler RC. The reliability of the functional independence measure: a quantitative review. Arch Phys Med Rehabil 1996; 77:1226-1232.
20. Hobart J, Freeman J, Lamping D, Fitzpatrick R, Thompson A. The SF-36 in multiple sclerosis: why basic assumptions must be tested. J Neurol Neurosurg Psychiatry 2001; 71:363-370.
21. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 1986; 42:121-130.
22. Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Qual Life Res 2003; 12:349-362.
23. Nortvedt MW, Riise T, Myhr KM, Nyland HI. Quality of life in multiple sclerosis: measuring the disease effects more broadly. Neurology 1999; 53:1098-1103.
24. Prins M, Geskus RB, Coutinho RA. [Roaming through methodology. XXXV. Bias in prevalent cohort studies of the natural course of disease]. Ned Tijdschr Geneeskd 2001; 145:2170-2172.
25. Alcabes P, Pezzotti P, Phillips AN, Rezza G, Vlahov D. Long-term perspective on the prevalent-cohort biases in studies of human immunodeficiency virus progression. Am J Epidemiol 1997; 146:543-551.

5

Vitality, perceived social support and disease activity determine social functioning in recently diagnosed multiple sclerosis: a longitudinal analysis

Vincent de Groot, Heleen Beckerman, Jos W.R. Twisk, Bernard M.J. Uitdehaag, Rogier Q. Hintzen, Arjan Minneboo, Gustaaf J. Lankhorst, Chris H. Polman & Lex M. Bouter

Submitted

Abstract

Objectives. To identify the strongest determinants of social functioning in multiple sclerosis (MS).

Method. A cohort of 156 recently diagnosed MS patients was prospectively followed for five measurements in three years. Social functioning was measured using the three social subscales of the Medical Outcome Study Short Form 36 (SF36) and a composite outcome based on these subscales. Potential determinants (n=43) were divided into the following clusters: patient and disease characteristics (n=12), psychosocial characteristics (n=10), basic functions (n=18), and basic activities (n=3). Results were analyzed with generalized estimating equations (GEE), using the following steps: 1) a stepwise backwards selection procedure for all clusters per outcome, 2) an overall stepwise backwards selection procedure for each outcome using the significant variables identified in step one, 3) examine whether the associations are based on within subject changes, *i.e.* is a change in the determinant for a patient associated with a change in the outcome for that same patient, 4) A sensitivity analysis using a p-value of 0.1.

Results. 1) In total 17 determinants were selected in any of the four models. 2) Vitality, the number of self-reported exacerbations and the perceived amount of social support were associated with social functioning in three or four of the models. 3) Almost all associations are based on within subject changes. 4) In the sensitivity analysis, additionally, the T2-weighted supratentorial lesion load was selected.

Conclusions. Vitality, the perceived amount of social support, and disease activity, *i.e.* the number of self-reported exacerbations and the T2-weighted supratentorial lesion load, determine social functioning. These results suggest that, in the early stages of MS, it might be beneficial to improve vitality, *e.g.* by promoting an active life-style, to optimize the perceived amount of social support, *e.g.* by counselling, and to suppress disease activity, *e.g.* by prescribing interferon.

Introduction

Multiple sclerosis (MS) is characterized by variable neurological symptomatology, which differs not only between patients but also within patients over time. This makes it difficult to predict the clinical course of the disease, which poses an important problem for clinicians treating MS patients. Reviews of the studies that looked at determinants of the clinical course showed that a progressive disease course, higher age at time of diagnosis, less than one year between relapses, and impairments of pyramidal or cerebellar tracts are associated with an unfavourable disease course, whereas an exacerbation as first sign of MS, a high recovery rate after the first exacerbation and afferent or monoregional symptoms are associated with a more favourable disease course.[1-7] Most studies focussed on neurological and locomotor function using the Expanded Disability Status Scale (EDSS) as outcome and the neurological deficits or MRI parameters as determinants.

Besides the neurological aspects of the clinical course, social functioning is important.[8,9] Longitudinal studies that looked at the social consequences of MS are rare.[10-12] In the first three years after the diagnosis MS has been made, when for most patients mobility and mental health are relatively mildly affected, about 40% of the patients report considerable limitations in social functioning.[13] For the whole group this percentage did not change in the first 3 years of the disease, but individual changes in social functioning did occur.

The aim of this study was to identify the strongest determinants that are longitudinally associated with social functioning in the first 3 years after the diagnosis MS has been made.

Methods

Patients and design

All consecutive potentially eligible patients visiting the participating outpatient clinics of neurology departments were asked to participate. A cohort of 156 recently (less than 6 months ago) diag-

nosed patients aged 16 through 55 years was recruited (in 1998-2000) and prospectively followed for 3 years. Diagnosis was made according to the Poser-criteria for definite MS.[14] Patients with other neurological disorders, systemic or malignant neoplastic diseases were excluded. Measurements took place at baseline, after 6 months, and after 1, 2 and 3 years. In case of a relapse, measurements were a few weeks postponed until the relapse had subsided. Patients were visited at home in order to minimise dropouts. Four well-trained raters performed the scoring.

Outcome measures

We used the Medical Outcome Study Short Form 36 (SF36)[15,16] subscales role physical (SF36rp), role emotional (SF36re) and social functioning (SF36sf) to measure the outcome social functioning. Validity and reliability have been extensively studied and found to be good.[15,16] Because of serious floor and ceiling effects, scores on all sub-scales were dichotomized using the mean -1.96 SD of an age-matched Dutch reference population[16] as cut-off. Sub-scale scores higher than this cut-off indicate normal social functioning, and, consequently, lower sub-scale scores indicate aberrant social functioning. Because all items of the 3 sub-scales are related to social functioning, we intended to create 1 summary social functioning score. For the most obvious solution, namely the summation of the 3 sub-scale scores, no support was found in the literature. Therefore, we created a composite social functioning outcome variable (SoFu) on the basis of the 3 dichotomized SF36 sub-scales. SoFu was normal, when all 3 SF36 sub-scale scores were normal, and SoFu was aberrant, when at least 1 of the 3 SF36 sub-scale scores was aberrant.

Determinants

Forty-three potential determinants, divided over four clusters that were based on the International Classification of Functioning,[17]

were identified using a literature search. Almost all determinants were measured on each point in time. Magnetic Resonance Imaging (MRI) data were only available at the baseline measurement. Scores on the questionnaires and cognitive tests were linearly transformed into a 0 to 10 scale. The cluster 'patient and disease characteristics' contains the following determinants: age (per 10 years), gender, comorbidity measured with the Cumulative Illness Rating Scale (CIRS, range 0 = no comorbidity to 10 maximal comorbidity score),[18,19] whether the disease starts with an exacerbation (non-relapse onset vs. relapse onset), the self-reported number of exacerbations in the previous period, time since first symptoms (logarithmically transformed), first neurological symptom (pyramidal, cerebellar, brainstem, sensory, bowel or bladder, optical; analyzed as five dummy variables), T1-hypointense and T2-weighted (supra- and infratentorial) lesion loads in cm^3 (MRI),[20] and number of lesions in the spinal cord (MRI).[21]

The cluster 'psychosocial characteristics' contains 10 determinants. Locus of control is measured with the sub-scales internal, physician and chance of the Multidimensional Health Locus of Control Scale;[22,23] sub-scale scores range from 0 = lowest to 10 = highest locus of control score. Personality traits were measured with the sub-scales psychoticism, extraversion and neuroticism of the Eysenck Personality Questionnaire;[24] sub-scale scores range from 0 = lowest to 10 = highest score on the personality trait. Two methods to measure social support were used. The *amount* of social support, *i.e.* a measure of *the quantity of supportive interactions*, was measured with the Social Support List Interactions;[25] scores range from 0 = no to 10 = maximal social support. The *perceived amount* of social support, *i.e.* a measure of *the extent to which the available supportive interactions cover the patient's need for social support*, was measured with the Social Support List Discrepancies;[25] scores range from 0 = needs are not covered at all to 10 = needs are completely covered. We also assessed whether the patient had a partner, and whether the patient had children.

The cluster 'basic functions' consists of 18 determinants. The involvement of the different neurological systems was measured with the Functional System (FS) scores, ranging from 0 = no impairment to 5 or 6 = maximal impairment) of the EDSS:[26] the FS optical, the FS brainstem, the FS cerebellar, the FS bowel and bladder, the FS pyramidal, the FS sensory, the FS mental and the FS other. Cognitive function was measured with the Brief Repeatable Battery of cognitive tests for MS,[27,28] which includes the sub-scales Consistent Long Term Retrieval and Long Term Storage of the Selective Reminding Test measuring verbal learning and memory, the 10/36 Spatial Recall Test measuring visuospatial learning and delayed recall, the Symbol Digit Modalities Test measuring sustained attention and concentration, the Paced Auditory Serial Addition Test measuring sustained attention and information processing speed, and the Word List Generation measuring verbal fluency; the scores range from 0 = worst possible to 10 best possible score. Fatigue was measured with the Fatigue Severity Scale,[29] which measures the patient's perceived level of fatigue in a variety of situations; scores range from 0 = lowest possible to 10 = highest possible fatigue score. Vitality, *i.e.* the presence of energy and the absence of fatigue, was measured with the SF36 sub-scale vitality;[15,16] scores range from 0 = lowest to 10 = maximal vitality score. Pain was assessed with the SF36 sub-scale bodily pain;[15,16] scores range from 0 = minimal to 10 = maximal pain score.

Finally, a cluster 'basic abilities' was created that consists of 3 determinants. Dexterity was measured with the Action Research Arm test,[30-32] and the Nine Hole Peg Test.[33-36] Ambulation was assessed with the 10-meter Timed Walk Test.[37]

Analysis

Statistical analyses were performed using *binomial* generalised estimating equations (GEE)[38] from the Statistical Package for Interactive Data Analysis (SPIDA) version 6.05 from the Statistical

Computing Laboratory. Time was modelled as a continuous variable expressed in years in every regression model. The analysis was performed in four steps:

1) A backwards selection procedure for every cluster per outcome with a significance level of 0.05 was used to reduce the large number of potential determinants.
2) The significant determinants identified in step 1 were entered into an overall regression model, and reduced using a backwards selection procedure with a significance level of 0.05. Results of the final regression models are presented as odds ratios (OR).
3) Because these OR contain information about the *between subject differences* (cross-sectional relationship) as well as information about the *within subject changes* (longitudinal relationship, *i.e.* is a change in the determinant for a patient associated with a change in the outcome for that same patient),[39] autoregressive models were created to disentangle the relative contribution of the *between subject differences* and the *within subject changes* to the OR. For this purpose, the outcome of the previous measurement was added as determinant to the regression models obtained in step 2. When the OR of the standard and autoregressive models are roughly similar, the effect can be attributed to within subject changes (longitudinal relationship). When the OR are closer to 1 (no effect) in the autoregressive model as compared with the standard model, the effect can be attributed to between subject differences (cross-sectional relationship). Thus, eight regression models were created (a standard and an autoregressive model for four outcome measures).
4) Because statistical modelling in small data sets is susceptible to bias,[40] we also performed a sensitivity analysis by repeating all steps in the analyses with a more liberal p-value of 0.10.

TABLE 5.1 Baseline characteristics of patients with multiple sclerosis.

	RR	SP	PP	Not yet known	Total
n (%)	120 (77%)	8 (5%)	25 (16%)	3 (2%)	156 (100%)
Age (SD)	35.5 (8.9)	48.2 (6.7)	43.2 (8.9)	45.5 (6.9)	37.6 (9.5)
Female (%)	84 (70.0%)	3 (37.5%)	11 (44%)	3 (100%)	101 (64%)
Years since diagnosis	0.26 (0.15-0.41)	0.33 (0.24-0.48)	0.28 (0.15-0.33)	0.14 (0.14-0.17)	0.26 (0.15-0.40)
Years since symptoms	1.83 (0.67-4.40)	7.50 (3.35-14.51)	2.10 (1.07-3.15)	3.62 (3.53-4.63)	2.15 (0.79-4.36)
Number of exacerbations	2 (1-3)	2 (1-7)	0 (0-0)	0 (0-0)	2 (1-3)
EDSS	2.0 (2.0-3.0)	3.0 (2.5-3.9)	3.0 (2.5-4.0)	2.5 (2.0-4.0)	2.5 (2.0-3.0)

n (percentage), mean (SD) or median (IQR). RR = relapsing remitting multiple sclerosis; SP = secondary progressive multiple sclerosis; PP = primary progressive multiple sclerosis. EDSS = Expanded Disability Status Scale, original score.

Results

Patients

TABLE 5.1 shows the baseline characteristics of the patients. Most characteristics comply with the expected pattern: more women than men in the RO group, more men than women in the NRO group, and more severe neurological deficits in the NRO group. Seven patients were lost to follow-up (3 after 1 year, 1 after 2 years and 3 after 3 years). Of the 149 patients with a complete follow-up 15 measurements were missing.

Reduction of determinants

In the first step of the analysis, where we analyzed the determinants for every cluster per outcome measure separately (using a p-value < 0.05), the list of 43 potential determinants was reduced to a list of 8 to 11 determinants for each outcome (TABLE 5.2). In total, 17 determinants were associated with one or more of the outcome measures.

Construction of standard models

Results of the second step in the analysis, the construction of the

TABLE 5.2 Candidate determinants for final regression models.

Determinants per cluster	SF36rp	SF36re	SF36sf	SoFu
Patient and disease characteristics				
Comorbidity	+		+	+
Exacerbations	+		+	+
Gender		*		
MRI: T2-weighted supratentorial lesion load	*	+		+
Psychosocial characteristics				
Locus of control: Internal			+	+
Locus of control: Physician		+		
Personality trait: Neuroticism	+	+	+	+
Personality trait: Psychoticism		+		
Social support: Amount of social support	+			+
Social support: Perceived social support	+	*	+	+
Basic functions				
Cognition: Sustained attention and concentration				*
Cognition: Verbal learning and memory: CLTR		*		
Cognition: Verbal learning and memory: LTS		*		
Cognition: Visuospatial learning delayed recall		+		
Fatigue	+			+
Functional systems: Cerebellar			+	
Functional systems: Mental	*	+	*	+
Functional systems: Optical		+		
Functional systems: Pyramidal			*	
Pain	*			
Vitality	+	+	+	+
Basic abilities				
Dexterity: ARA			+	
Dexterity: NHPT	+	+		+

SF36rp: Medical Outcome Study Short Form 36 sub-scale role physical; SF36re: SF36 sub-scale role emotional; SF36sf: SF36 sub-scale social functioning; SoFu: composite score based on the SF36rp, SF36re and SF36sf; CLTR: Consistent Long Term Retrieval; LTS: Long Term Storage; ARA: Action Research Arm Test; NHPT: Nine Hole Peg Test. + p-value < 0.05 in stepwise backwards selection procedure per cluster per outcome. * p-value <0.10 in stepwise backwards selection procedure per cluster per outcome (sensitivity analysis).

standard models, can be found in TABLE 5.3. Vitality, measured with the SF36vt, is associated with all outcome measures (OR 1.73-2.41), indicating that patients reporting more vitality have higher odds (the ratio of the probability that the patient does not

TABLE 5.3 Determinants of social functioning of recently diagnosed multiple sclerosis patients.

	SF36rp		SF36re	
Determinant	standard	autoregression	standard	autoregression
Vitality*	1.80 (1.57-2.07)	1.88 (1.61-2.20)	1.73 (1.37-2.19)	1.80 (1.37-2.37)
Perceived social support*	1.21 (1.05-1.39)	1.08 (0.91-1.29)		
Exacerbations†	0.80 (0.68-0.95)	0.66 (0.47-0.92)		
Amount of social support*	0.75 (0.62-0.91)	0.79 (0.63-1.01)		
Fatigue*	0.82 (0.73-0.92)	0.87 (0.76-1.00)		
Visuospatial learning delayed recall*			1.28 (1.10-1.50)	1.36 (1.14-1.63)
Psychoticism*			0.64 (0.48-0.84)	0.62 (0.44-0.86)
Neuroticism*			0.77 (0.69-0.87)	0.79 (0.69-0.91)
Locus of Control physician*				1.23 (0.99-1.53)
Comorbidity*				
T2-weighted supratentorial lesion load (cm^3)				

Odds ratios (95%CI) for standard and autoregression models. Outcomes are dichotomized using the mean -1.96 SD of a healthy Dutch reference population; for all outcomes 1= normal social functioning and 0=aberrant social functioning. SF36rp =

deviate from the norm to the probability that he does) to have normal social functioning. Besides this being a consistent association, it is also the strongest association among the determinants studied.

The number of self reported exacerbations (OR 0.77-0.88) and the *perceived amount* of social support (OR 1.17-1.34), *i.e.* the extent to which the available supportive interactions cover the patient's need for social support, are associated with three outcome measures and, indicating that patients who report more exacerbations have lower and that patients who perceive more social support have higher odds to have normal social functioning.

Fatigue (OR 0.82-0.84) and the *amount* of social support (OR 0.71-0.75), *i.e.* the quantity of supportive interactions, are associated with two outcome measures, indicating that more fatigued patients and patients having a large amount of social support have lower odds to have normal social functioning. The other determi-

nts.

SF36sf		SoFu	
standard	autoregression	standard	autoregression
2.41 (1.94-2.99)	2.29 (1.85-2.85)	2.12 (1.81-2.48)	2.32 (1.94-2.76)
1.34 (1.19-1.50)	1.23 (1.08-1.42)	1.17 (1.02-1.35)	1.13 (0.96-1.32)
0.83 (0.69-1.00)	0.59 (0.42-0.83)	0.77 (0.64-0.93)	0.64 (0.45-0.89)
		0.71 (0.59-0.87)	0.74 (0.60-0.92)
		0.84 (0.73-0.96)	0.91 (0.80-1.05)
1.28 (1.01-1.62)			
0.25 (0.09-0.67)	0.19 (0.04-0.86)		
		0.74 (0.59-0.94)	0.74 (0.60-0.92)

Medical Outcome Study Short Form 36 sub-scale Role Physical; SF36re = SF36 sub-scale Role Emotional; SF36sf = SF36 sub-scale Social Functioning; SoFu = composite score based on the SF36rp, SF36re and SF36sf. * Scaled from 0 to 10. † number.

nants are associated with only one outcome measure. Most of these are associated with the SF36 sub-scale role emotional.

Construction of autoregressive models

In step 3, the construction of the autoregressive models, most odds ratios are similar to the odds ratios in the standard models, indicating that the association can be attributed to *within subject changes* (longitudinal relationship). However, the association of the perceived amount of social support with the SF36 sub-scale role physical is somewhat reduced, indicating that this association is based on both between subject differences (cross-sectional relationship) and within subject changes (longitudinal relationship).

Sensitivity analysis

In step 4, the sensitivity analyses (using a p-value < 0.10), five additional variables were selected (TABLE 5.2). In the subsequently con-

structed standard models, visuospatial learning delayed recall and comorbidity were no longer included in any model, the T2-weighted supratentorial lesion load was added to the SF36 sub-scale role physical [OR 0.78 (95%CI 0.62-0.99)] and SF36 sub-scale role emotional model [OR 0.71 (95%CI 0.57-0.88)], verbal learning and memory (consistent long term retrieval) and gender were added to the SF36 sub-scale role emotional model [OR 1.26 (95%CI 1.01-1.56) and 0.16 (95%CI 0.06-0.44)], and FS cerebellar was added to the SF36 sub-scale social functioning model [OR 0.70 (95%CI 0.51-0.97)]. The OR did not change considerably in the autoregression models, which means that associations are based on within subject changes (longitudinal relationship).

Discussion

The most important determinants associated with social functioning are vitality and the *perceived amount* of social support. Other relevant determinants are the number of self-reported exacerbations and the T2-weighted supratentorial lesion load, although these are less consistently associated with social functioning. The associations are based on within subject changes, which means that a change in the determinant for a particular patient is associated with a change in the outcome for that same patient.

Until now, studies that looked for determinants associated with social functioning have predominantly been cross-sectional.[41-45] An important strength of our study is its longitudinal design and analysis, and the simultaneous assessment of an extensive set of determinants and outcome measures. This enables a direct comparison of the adjusted association of several determinants with the outcome measures. Furthermore, it is possible to use longitudinal data analysis techniques, which enable investigation of the contribution of between subject differences (cross-sectional relationship) and within subject changes (longitudinal relationship) to the regression coefficients.

A possible weakness is the use of multiple raters. Due to the 6-year study period four raters had to be trained to perform the measurement, probably at the risk of reducing reliability of the measurements. For the measures that require skilled raters, such as the cognitive tests and the scoring of the functional systems of the EDSS, an intensive training was given, and new raters were supervised until they were adequately able to perform the measurements.

Another possible weakness is the relatively small sample size of 156 patients. Modelling in small samples might influence the selection of variables. Simulation experiments have shown that stepwise methods have limited power to select important determinants, and, on the other hand, carry the risk that one or more (almost) random determinants are selected, since multiple comparisons are made.[40] Furthermore, the estimate of the regression coefficient might be biased.[40] In order to minimise these effects, we used longitudinal regression analysis (GEE) to describe the longitudinal relationships. GEE has the important advantage that all available data are used, which increases the power to detect subtle relationships. Furthermore, we also performed a sensitivity analysis with a more liberal p-value of 0.10. This resulted in the exclusion of visuospatial learning delayed recall and comorbidity from the models in which they were previously included, and the addition of one or three variables for the outcomes SF36 sub-scales role physical, role emotional and social functioning.

This study shows some consistent findings across the different aspects of social functioning, which suggests that the determinants vitality, the *perceived amount* of social support, the number of self-reported exacerbations and the T2-weighted supratentorial lesion load might play a role in the development of social dysfunction. However, the results of this study should be interpreted with caution, because causal inferences cannot be made. From a clinical point of view it can easily be understood that diminished vitality, i.*e.* a lack of energy, leads to social dysfunction. Patients might have

problems to continue sports or leisure activities, skip social activities with family or friends, or are no longer able to work a full day. Also, it is logical to assume that an increase in the number of self-reported exacerbations will lead to social dysfunction. Although the relationship between an increase in T2-weighted supratentorial lesion load and more social dysfunction is also imaginable, the precise mechanism is less obvious. A higher T2-weighted supratentorial lesion load might be an indicator of cognitive dysfunction,[46-49] or might indicate that there has been more disease activity (accumulated pathological changes). In our sample, there was no association of T2-weighted supratentorial lesion load with time since first symptoms, which indicates that T2-weighted lesion load is not an indicator of time since first symptoms.

The interpretation of the associations with the two social support scales is complex, and further complicated by the fact that the associations of both scales point in opposite directions. The correlation between both scales is low (rho=0.44), indicating that they measure different constructs, and contribute independently to the results of the analyses. The association with the *perceived amount* of social support can be interpreted in two ways. It might be that a patient, who experiences a lack of social support, is less inclined to participate in social activities. But, the relationship might also be interpreted in the opposite direction: a patient, who shows social dysfunction for any reason, experiences this as a lack of social support. The interpretation of the finding that patients who report a greater *amount* of social support have higher odds of social dysfunctioning should probably be that social dysfunctioning leads to an accumulation of supportive interactions, and thus an increase in the *amount* of social support. Finally, an overall interpretation of social support might be that a low *perceived amount* of social support leads to a compensatory increase in the *amount* of social support, which is, unfortunately, not sufficient to normalize social functioning.

Future studies should focus on further clarification of the possible causal relationship between vitality, the *perceived amount* of

social support, the number of self-reported exacerbations and the T2-weighted supratentorial lesion load, and social functioning. It would be very interesting to design trials that study the effect of interventions, developed to influence the above mentioned determinants, on social functioning. Trials showing that these interventions lead to improved social functioning would provide strong evidence for a causal relationship.

But first, more longitudinal studies are needed to build a broader evidence base. Our results might aid the selection of determinants for these studies. Since we realize that the 3-year follow-up of this cohort is relatively short, we will try to extend the follow-up. This would enable a description of the course of social functioning during the later stages of the disease and a study of the determinants associated with social functioning in these later stages. Patients in our incidence cohort are currently only relatively mildly disabled, which will certainly change with longer disease duration. This will probably have consequences for the relative importance of the determinants we have studied.

Clinicians caring for patients with MS will be confronted with patients that are limited in their social functioning. Interventions to reduce the number of exacerbations and the development of new lesions are available and implemented in clinical practice. Our results indicate that vitality and the *perceived amount* of social support might also be important factors to assess in these patients. Isolated interventions to improve vitality, such as amantadine,[50,51] energy conservation techniques[52,53] and aerobic training,[54] are available, but evidence regarding their efficacy is not conclusive. Different interventions to improve the *perceived amount* of social support are also available, but again the evidence is not conclusive.[55] Furthermore, as argued above, is has not been shown that these interventions lead to improved social functioning. However, it has been shown that outpatient as well as home based multidisciplinary rehabilitation programmes can lead to improvements in social functioning.[56-58] These studies looked at the rehabilitation

programme as a whole and did not focus on specific elements in these programmes. It is very likely, though, that these programmes contained elements addressing vitality and the perceived amount of social support.

Conclusions

Vitality, the perceived amount of social support, and disease activity, *i.e.* the number of self-reported exacerbations and the T2-weighted supratentorial lesion load, determine social functioning. These results suggest that, in the early stages of MS, it might be beneficial to improve vitality, *e.g.* by promoting an active life-style, to optimize the perceived amount of social support, *e.g.* by counselling, and to suppress disease activity, *e.g.* by prescribing interferon.

References

1. Runmarker B, Andersen O. Prognostic factors in a multiple sclerosis incidence cohort with twenty-five years of follow-up. Brain 1993; 116:117-134.
2. Weinshenker BG, Bass B, Rice GPA, Noseworthy J, Carriere W, Baskerville J *et al.* The natural history of multiple sclerosis: A geographically based study. 2. Predictive value of the early clinical course. Brain 1989; 112:1419-1428.
3. Weinshenker BG, Bass B, Rice GPA, Noseworthy J, Carriere W, Baskerville J *et al.* The natural history of multiple sclerosis: A geographically based study. I. Clinical course and disability. Brain 1989; 112:133-146.
4. Weinshenker BG, Rice GPA, Noseworthy JH, Carriere W, Baskerville J, Ebers GC. The natural history of multiple sclerosis: A geographically based study. 3. Multivariate analysis of predictive factors and models of outcome. Brain 1991; 114:1045-1056.
5. Amato MP, Ponziani G. A prospective study on the prognosis of multiple sclerosis. Neurol Sci 2000; 21:S831-S838.
6. Confavreux C, Vukusic S, Adeleine P. Early clinical predictors and progression of irreversible disability in multiple sclerosis: an amnesic process. Brain 2003; 126:770-782.
7. Pittock SJ, Mayr WT, McClelland RL, Jorgensen NW, Weigand SD, Noseworthy JH *et al.* Change in MS-related disability in a population-based cohort: a 10-year follow-up study. Neurology 2004; 62:51-59.

8. Murphy N, Confavreux C, Haas J, Konig N, Roullet E, Sailer M *et al.* Quality of life in multiple sclerosis in France, Germany, and the United Kingdom. J Neurol Neurosurg Psychiatry 1998; 65:460-466.
9. Rothwell PM. Quality of life in multiple sclerosis. J Neurol Neurosurg Psychiatry 1998; 65:433.
10. Morales-Gonzales JM, Benito-Leon J, Rivera-Navarro J, Mitchell AJ. A systematic approach to analyse health-related quality of life in multiple sclerosis: the GEDMA study. Mult Scler 2004; 10:47-54.
11. Di Fabio RP, Choi T, Soderberg J, Hansen CR. Health-related quality of life for patients with progressive multiple sclerosis: influence of rehabilitation. Phys Ther 1997; 77:1704-1716.
12. Haupts M, Elias G, Hardt C, Langenbahn H, Obert H, Pohlau D *et al.* [Quality of life in patients with remitting-relapsing multiple sclerosis in Germany]. Nervenarzt 2003; 74:144-150.
13. de Groot V, Beckerman H, Lankhorst GJ, Polman CH, Bouter LM. The initial course of daily functioning in multiple sclerosis: a three-year follow-up study. Mult Scler 2005; 11:713-718.
14. Poser CM, Paty DW, Scheinberg L, McDonald WI, Davis FA, Ebers GC *et al.* New diagnostic criteria for multiple sclerosis: guidelines for research protocols. Ann Neurol 1983; 13:227-231.
15. Freeman JA, Hobart JC, Langdon DW, Thompson AJ. Clinical appropriateness: a key factor in outcome measure selection: the 36 item short form health survey in multiple sclerosis. J Neurol Neurosurg Psychiatry 2000; 68:150-156.
16. Aaronson NK, Muller M, Cohen PD, Essink-Bot ML, Fekkes M, Sanderman R *et al.* Translation, validation, and norming of the Dutch language version of the SF-36 Health Survey in community and chronic disease populations. J Clin Epidemiol 1998; 51:1055-1068.
17. World Health Organization. International Classification of Functioning, Disability and Health. http://www3.who.int/icf/icftemplate.cfm, 2001. Geneva, World Health Organization.
18. Linn BS, Linn MW, Gurel L. Cumulative illness rating scale. Journal of the American Geriatrics Society 1968; 16:622-626.
19. de Groot V, Beckerman H, Lankhorst GJ, Bouter LM. How to measure comorbidity. a critical review of available methods. J Clin Epidemiol 2003; 56:221-229.
20. Kalkers NF, Bergers L, de Groot V, Lazeron RH, van Walderveen MA, Uitdehaag BM *et al.* Concurrent validity of the MS Functional Composite using MRI as a biological disease marker. Neurology 2001; 56:215-219.
21. Bot JC, Barkhof F, Polman CH, Nijeholt GJ, de Groot V, Bergers E *et al.* Spinal cord abnormalities in recently diagnosed MS patients: added value of spinal MRI

examination. Neurology 2004; 62:226-233.
22. Wallston KA, Wallston BS, DeVellis R. Development of the Multidimensional Health Locus of Control Scales. Health Educ Monogr 1978; 6:160-170.
23. Halfens RJ, Philipsen H. Een gezondheidsspecifieke beheersingsorientatie-schaal: validiteit en betrouwbaarheid van de MHLC. T Soc Gezondh 1988; 66:399-403.
24. Sanderman R, Arrindell WA, Ranchor AV, Eysenck HJ, Eysenck SBG. Eysenck Personality Questionnaire (EPQ). Groningen: Noordelijk Centrum voor Gezondheidsvraagstukken Rijksuniversiteit Groningen, 1997.
25. van Sonderen E. Sociale Steun Lijst – Interakties (SSL-I) en Sociale Steun Lijst – Discrepanties (SSL-D). Groningen: Noordelijk Centrum voor Gezondheids-vraagstukken Rijksuniversiteit Groningen, 1997.
26. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology 1983; 33:1444-1452.
27. Rao SM, Leo GJ, Bernardin L, Unverzagt F. Cognitive dysfunction in multiple sclerosis. I. Frequency, patterns, and prediction. Neurology 1991; 41:685-691.
28. Boringa JB, Lazeron RH, Reuling IE, Adèr HJ, Pfennings L, Lindeboom J *et al.* The brief repeatable battery of neuropsychological tests: normative values allow application in multiple sclerosis clinical practice. Mult Scler 2001; 7:263-267.
29. Krupp LB, LaRocca NG, Muir-Nash J, Steinberg AD. The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythe-matosus. Arch Neurol 1989; 46:1121-1123.
30. Lyle RC. A performance test for assessment of upper limb function in physi-cal rehabilitation treatment and research. Int J Rehabil Res 1981; 4:483-492.
31. van der Lee JH, Beckerman H, Lankhorst GJ, Bouter LM. The responsiveness of the Action Research Arm test and the Fugl-Meyer Assessment scale in chron-ic stroke patients. J Rehabil Med 2001; 33:110-113.
32. van der Lee JH, de Groot V, Beckerman H, Wagenaar RC, Lankhorst GJ, Bouter LM. The intra- and interrater reliability of the action research arm test: a practical test of upper extremity function in patients with stroke. Arch Phys Med Rehabil 2001; 82:14-19.
33. Goodkin DE, Hertsgaard D, Seminary J. Upper extremity function in multiple sclerosis: improving assessment sensitivity with box-and-block and nine-hole peg tests. Arch Phys Med Rehabil 1988; 69:850-854.
34. Cohen JA, Fischer JS, Bolibrush DM, Jak AJ, Kniker JE, Mertz LA *et al.* Intrarater and interrater reliability of the MS functional composite outcome measure. Neurology 2000; 54:802-806.
35. Fischer JS, Rudick RA, Cutter GR, Reingold SC. The Multiple Sclerosis Functional Composite Measure (MSFC): an integrated approach to MS clinical

outcome assessment. National MS Society Clinical Outcomes Assessment Task Force. Mult Scler 1999; 5:244-250.

36. Cutter GR, Baier ML, Rudick RA, Cookfair DL, Fischer JS, Petkau J *et al.* Development of a multiple sclerosis functional composite as a clinical trial outcome measure. Brain 1999; 122:871-882.
37. Rossier P, Wade DT. Validity and reliability comparison of 4 mobility measures in patients presenting with neurologic impairment. Arch Phys Med Rehabil 2001; 82:9-13.
38. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 1986; 42:121-130.
39. Twisk JW. Applied Longitudinal Data Analysis for Epidemiology: a Practical Guide. 1 ed. Cambridge: Cambridge University Press, 2003.
40. Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol 1999; 52:935-942.
41. Hemmett L, Holmes J, Barnes M, Russell N. What drives quality of life in multiple sclerosis? QJM 2004; 97:671-676.
42. Pittock SJ, Mayr WT, McClelland RL, Jorgensen NW, Weigand SD, Noseworthy JH *et al.* Disability profile of MS did not change over 10 years in a population-based prevalence cohort. Neurology 2004; 62:601-606.
43. Brunet DG, Hopman WM, Singer MA, Edgar CM, MacKenzie TA. Measurement of health-related quality of life in multiple sclerosis patients. Can J Neurol Sci 1996; 23:99-103.
44. Somerset M, Peters TJ, Sharp DJ, Campbell R. Factors that contribute to quality of life outcomes prioritised by people with multiple sclerosis. Qual Life Res 2003; 12:21-29.
45. Somerset M, Sharp D, Campbell R. Multiple sclerosis and quality of life: a qualitative investigation. J Health Serv Res Policy 2002; 7:151-159.
46. Lazeron RH, Boringa JB, Schouten M, Uitdehaag BM, Bergers E, Lindeboom J *et al.* Brain atrophy and lesion load as explaining parameters for cognitive impairment in multiple sclerosis. Mult Scler 2005; 11:524-531.
47. Rovaris M, Iannucci G, Falautano M, Possa F, Martinelli V, Comi G *et al.* Cognitive dysfunction in patients with mildly disabling relapsing-remitting multiple sclerosis: an exploratory study with diffusion tensor MR imaging. J Neurol Sci 2002; 195:103-109.
48. Camp SJ, Stevenson VL, Thompson AJ, Miller DH, Borras C, Auriacombe S *et al.* Cognitive function in primary progressive and transitional progressive multiple sclerosis: a controlled study with MRI correlates. Brain 1999; 122:1341-1348.

49. Miller DH, Grossman RI, Reingold SC, Mcfarland HF. The role of magnetic resonance techniques in understanding and managing multiple sclerosis. Brain 1998; 121:3-24.
50. Krupp LB, Christodoulou C. Fatigue in multiple sclerosis. Curr Neurol Neurosci Rep 2001; 1:294-298.
51. Krupp LB, Coyle PK, Doscher C, Miller A, Cross AH, Jandorf L *et al.* Fatigue therapy in multiple sclerosis: results of a double-blind, randomized, parallel trial of amantadine, pemoline, and placebo. Neurology 1995; 45:1956-1961.
52. Mathiowetz VG, Matuska KM, Murphy ME. Efficacy of an energy conservation course for persons with multiple sclerosis. Arch Phys Med Rehabil 2001; 82:449-456.
53. Mathiowetz VG, Finlayson ML, Matuska KM, Chen HY, Luo P. Randomized controlled trial of an energy conservation course for persons with multiple sclerosis. Mult Scler 2005; 11:592-601.
54. Petajan JH, Gappmaier E, White AT, Spencer MK, Mino L, Hicks RW. Impact of aerobic training on fitness and quality of life in multiple sclerosis. Ann Neurol 1996; 39:432-441.
55. Hogan BE, Linden W, Najarian B. Social support interventions – Do they work? Clin Psychol Rev 2002; 22:381-440.
56. Freeman JA, Langdon DW, Hobart JC, Thompson AJ. The impact of inpatient rehabilitation on progressive multiple sclerosis. Ann Neurol 1997; 42:236-244.
57. Patti F, Ciancio MR, Reggio E, Lopes R, Palermo F, Cacopardo M *et al.* The impact of outpatient rehabilitation on quality of life in multiple sclerosis. J Neurol 2002; 249:1027-1033.
58. Pozzilli C, Brunetti M, Amicosante AM, Gasperini C, Ristori G, Palmisano L *et al.* Home based management in multiple sclerosis: results of a randomised controlled trial. J Neurol Neurosurg Psychiatry 2002; 73:250-255.

6

Physical and cognitive functioning can be predicted in recently diagnosed multiple sclerosis

Vincent de Groot, Heleen Beckerman, Bernard M.J. Uitdehaag, Rogier Q. Hintzen, Arjan Minneboo, Martijn W. Heymans, Gustaaf J. Lankhorst, Chris H. Polman & Lex M. Bouter

Submitted

Abstract

Objectives. To predict functioning after three years in patients with recently diagnosed multiple sclerosis (MS).

Methods. 146 recently diagnosed MS patients were monitored for three years. At baseline, predictors were obtained from history-taking, neurological examination and Magnetic Resonance Imaging. The outcomes of interest after three years were: inability to walk at least 500 metres, impaired dexterity, cognitive impairments, incontinence, inability to drive a car or use public transportation, social dysfunction, and reliance on a disability pension. Clinical prediction rules were constructed for the models that were well calibrated (sufficient agreement between predicted and observed outcomes, based on visual inspection of calibration curves) and that showed sufficient discrimination (Area under the Receiver Operation Characteristic Curve > 0.70) after internal bootstrap validation.

Results. The models for inability to walk at least 500 metres, impaired dexterity and cognitive impairments were well calibrated. Discrimination was sufficient for all seven models, except the one predicting social dysfunction (0.67).

Conclusions. Inability to walk at least 500 metres, impaired dexterity and cognitive impairments can be predicted with predictors that are obtained shortly after the definite diagnosis has been made. The ability to predict physical and cognitive functioning might facilitate the counselling of patients with MS and the planning of their (rehabilitation) treatment.

Introduction

Multiple sclerosis (MS) is characterized by variable neurological symptomatology, which differs not only between patients but also within patients over time. This makes it difficult to predict the clinical course of the disease, and therefore it is an important problem for clinicians treating MS patients. Furthermore, it causes uncertainty about the future for patients with MS, and this has a negative influence on their quality of life.[1,2] Well validated prognostic models can aid clinicians in making decisions about certain (preventive) treatments for patients with MS or can improve the information given to these patients about their future prognosis.

Thus far, the prediction models published in the literature on MS had a strong focus on the strength and the relevance of the predictors themselves,[3-10] hoping that this would provide clues to better understanding of the aetiology of the disease or the course of the disease. Reviews of the studies that have investigated determinants of the course of the disease showed that a progressive onset, higher age at the time of diagnosis, less than one year between relapses, and impairments of pyramidal or cerebellar tracts are associated with a progressive disease course, whereas an exacerbation as first sign of MS, a high recovery rate after the first exacerbation and afferent or monoregional symptoms are associated with a more favourable disease course.[3-10] In contrast with the literature on cardiologic disorders and intensive care units,[11,12] the literature on MS has not yet assessed the usefulness of the complete prognostic models to predict future events accurately. The construction of these prediction models differs with regard to the selection of variables, the way of fitting the models, and the presentation of the results.

With respect to future outcomes, most studies have focussed on neurological and locomotor function, using the score of the Expanded Disability Status Scale (EDSS) as outcome and the neurological deficits or Magnetic Resonance Imaging (MRI) parameters as candidate predictors. However, there are other areas of func-

tioning that are relevant for patients, such as wheelchair-dependence, impaired dexterity, cognitive impairments, incontinence, inability to use a car or public transportation, social dysfunction and reliance on a disability pension. Studying these outcomes also means that the predictors should not be limited to neurological or MRI parameters, but that psychosocial predictors should also be assessed.

The aim of our study was to construct and assess the usefulness of prediction models to predict functioning in the areas of mobility, dexterity, cognition, voiding, transportation, social activities and work.

Methods

Patients and design

All consecutive potentially eligible patients visiting the participating outpatient clinics of five neurology departments were invited to participate. A cohort of 156 recently (less than six months previously) diagnosed patients, aged 16-55 years, was recruited in 1998-2000, and prospectively monitored for three years. Diagnosis was determined according to the Poser-criteria for definite MS.[13] Patients with other neurological disorders, systemic diseased or malignant neoplastic diseases were excluded. This study was performed as part of a longitudinal study collecting extensive data on many potentially relevant predictors and outcomes at baseline, and at six months, and one, two and three years later.[14] For the present analyses we used the baseline information for the predictors, and the three-year data for the relevant outcomes. The patients were visited at home in order to minimise drop-out, and four well-trained raters were responsible for the scoring.

Construction of prediction models

The prediction models were constructed with the intention to use them in clinical practice. Therefore, we involved representatives of potential users of these models in the construction phase. Prior to

actual data-analysis, the aims of our study were discussed during two informal semi-structured workshops with neurologists and researchers specialized in MS, and with rehabilitation physicians, and physical and occupational therapists. In these workshops, we discussed which outcomes would be relevant to predict, and which candidate predictors should be investigated to predict these outcomes.

Outcomes

Inability to walk at least 500 metres was defined as an EDSS score ≥ 4.[15] Impaired dexterity was defined as an abnormal score (mean – 1.96 x SD, healthy Dutch reference population) for the Nine Hole Peg Test.[16] Cognitive impairments was defined as a score of mean – 1 x SD for one or more sub-tests of a cognitive screening test that was specifically developed for MS.[17-19] Incontinence was defined as a score ≤ 5 for the continence item of the Functional Independence Measure.[20] Inability to drive a car or use public transportation was defined as needing help or being unable on the ability to travel item of the Rehabilitation Activities Profile.[21] Social dysfunction was defined as an abnormal score (mean – 1.96 x SD, healthy Dutch reference population) for one or more of the three social sub-scales (role physical, role emotional, social functioning) of the Medical Outcome Study Short Form 36.[22] The patients were asked in a direct question about complete or partial reliance on a disability pension.

Candidate predictors

Participants in the workshops were encouraged to name predictors that are relatively easy to obtain in clinical practice. First of all, the most relevant predictors for which information could be gathered during medical history-taking were identified. Next, the most relevant predictors for which a physical examination is required were identified, and finally, the most relevant predictors obtained through complex diagnostic tests were identified. Using the information obtained from the discussions and from the literature, as

TABLE 6.1 Candidate predictors measured at baseline for each outcome of interest.

Predictor per outcome of interest	Range	Description
Inability to walk at least 500 m		
Medical history-taking		
How well can you walk? [1]	0-10	Not at all – very well
Are you easily tired? [1]	0-10	Very easily – not at all
Physical examination		
Impairment of pyramidal tract [2]	0-6	No signs – quadriplegia
Impairment of cerebellar tract [2]	0-5	No signs – severe ataxia
MRI-parameter		
Number of lesions in spinal cord [3]	number	Number of lesions counted
Impaired dexterity		
Medical history-taking		
How well can you use your hands? [1]	0-10	Not at all – very well
Physical examination		
Impairment of sensory tract [2]	0-6	No signs – sensation lost below head
Impairment of pyramidal tract [2]	0-6	No signs – quadriplegia
Impairment of cerebellar tract [2]	0-5	No signs – severe ataxia
MRI-parameter		
T2-weighted infratentorial lesion load [3]	cm^3	
Cognitive impairments		
Medical history-taking		
Age	Years	
Gender	0-1	Female – male
How good is your memory? [1]	0-10	Bad – good
Can you concentrate? [1]	0-10	Not at all – very well
Physical examination		
none		
MRI-parameter		
T2-weighted supratentorial lesion load [3]	cm^3	
Incontinence		
Medical history-taking		
Can you contain your urine well? [1]	0-10	Not at all – easily
Physical examination		
Impairment of pyramidal tract [2]	0-6	No signs – quadriplegia
MRI-parameter		
Number of lesions in spinal cord [3]	number	Number of lesions counted

TABLE 6.1 - continued

Predictor per outcome of interest	Range	Description
Inability to use a car or public transportation		
Medical history-taking		
How good is your memory? [1]	0-10	Bad – good
Can you concentrate? [1]	0-10	Not at all – very well
Physical examination		
Impairment of pyramidal tract [2]	0-6	No signs – quadriplegia
Impairment of cerebellar tract [2]	0-5	No signs – severe ataxia
MRI-parameter		
T2-weighted total lesion load [3]	cm^3	
Number of lesions in spinal cord [3]	number	Number of lesions counted
Social dysfunction		
Medical history-taking		
How good is your contact with members of your household? [1]	0-10	Bad – excellent
How do you feel? [1]	0-10	Gloomy – happy
Are you easily tired? [1]	0-10	Very easily – not at all
Physical examination		
Impairment of pyramidal tract [2]	0-6	No signs – quadriplegia
Impairment of cerebellar tract [2]	0-5	No signs – severe ataxia
MRI-parameter		
T2-weighted total lesion load [3]	cm^3	
Reliance on a disability pension		
Medical history-taking		
How do you feel? [1]	0-10	Gloomy – happy
How good is your memory? [1]	0-10	Bad – good
Can you concentrate? [1]	0-10	Not at all – very well
Are you easily tired? [1]	0-10	Very easily – not at all
Physical examination		
Impairment of pyramidal tract [2]	0-6	No signs – quadriplegia
Impairment of cerebellar tract [2]	0-5	No signs – severe ataxia
MRI-parameter		
T2-weighted total lesion load [3]	cm^3	

[1] Item of the Disability and Impact Profile; [2] Item of the Functional Systems of the Expanded Disability Status Scale; [3] Values derived from Magnetic Resonance Imaging of the brain and spinal cord.

described in the introduction, we selected candidate predictors from the baseline data of the extensive data-set. For the predictors that are based on medical history-taking we used items of the Disability and Impact Profile.[23,24] For the predictors that are based on physical examination we used the EDSS Functional Systems scores.[15] MRI was used to obtain the predictor variables T2-weighted (supra- and infratentorial) lesion loads in cm^3, and the number of lesions in the spinal cord.[25,26] TABLE 6.1 shows the selected outcomes and the predictors that were used to construct the models.

Analysis

Only data including complete follow-up at three years were analyzed. Missing data on predictors were imputed twice, using the data-augmentation procedure in NORM software,[27] yielding two imputed data-sets. Descriptive statistics were used to describe the study population. For each outcome the number and percentage of patients with an unfavourable outcome was calculated.

Because predictive modelling in small data-sets is susceptible to bias, we made use of the approach described by Steyerberg *et al.*[28,29]. We used a limited set of candidate predictors that were selected on the basis of information from the literature and on clinical grounds. Subsequently, logistic regression models were constructed in each imputed data-set, using a backwards stepwise selection procedure with a liberal p-value of 0.5. When predictors in these models showed a contra-intuitive relationship with the outcome, which means that the sign of the regression coefficient is opposite to what we expected, this predictor was deleted from the model, and the backwards selection procedure was repeated. Because the selected predictors were the same in both imputed data-sets, internal validation was performed on one of the sets.

Bootstrapping techniques were used to study the internal validity of the final models, i.e. to adjust the estimated regression coefficients for over-fitting and the model performance for over-optimism.[29,30] Random bootstrap samples were drawn with

replacement (250 replications) from the full data-set. The shrinkage factor, a result of the bootstrap analyses, is a measure of over-fitting. Regression coefficients can be corrected for over-fitting by multiplying them by this shrinkage factor. Bootstrapping was performed in S-plus 6.1 (Insightful Corp., Seattle, WA, USA).

Model performance

The model performance, expressed as calibration and discrimination, after bootstrapping can be considered as the performance that can be expected from similar future patients. Calibration refers to whether the predicted outcomes agree with the observed outcomes. A frequently occurring problem with prediction models is that the predictions for new patients are too extreme (too high for high-risk patients and too low for low-risk patients). Well-calibrated models have a slope of 1, while models providing predictions that are too extreme have a slope of less than 1.

The discriminative ability of the model, i.e. how accurately can high-risk patients be distinguished from low-risk patients, was assessed using the Area Under the receiver operation characteristic curve (AUC; 95%CI). An AUC of 0.5 indicates no discrimination above chance, whereas an AUC of 1.0 indicates perfect discrimination. A rough guide for classifying the discriminative ability of a diagnostic test is the traditional academic points system: excellent (>0.9), good (>0.8), fair (>0.70), poor (>0.60) or fail (>0.50).[31]

Clinical prediction rules

To facilitate the calculation of an individual patient's risk, we developed score charts for the prediction models that were internally valid. We divided the regression coefficients of the multivariate models by the lowest regression coefficient and rounded them to the nearest integer to form scores for the predictors. The sum of the scores corresponds to the risk of a poor outcome.

Results

Patients

Data on the outcomes at the three-year follow-up were missing for 10 of the 156 patients. These 10 patients did not differ significantly from the rest of the cohort with regard to gender, age, T2-weighted lesion load at baseline, or number of lesions in the spinal cord at baseline. However, they had a trend towards higher baseline EDSS scores, and, contradictory to the results for the EDSS, less lesions on the baseline MRI. For 13 of the 146 patients with a complete follow-up, baseline MRI data on brain and spinal cord were missing. MRI data on the spinal cord were also missing for two patients. These data were imputed. Data on all other candidate predictors were complete. TABLE 6.2 shows the baseline character-

TABLE 6.2 Baseline characteristics of the 146 MS patients.

Patient characteristics	
Female	93 (64%)
Age	37.4 (SD 9.7)
Disease characteristics	
Relapse Onset (RO)	120 (82%)
Non-relapse Onset (NRO)	26 (18%)
Expanded Disability Status Scale	2.5 (IQR 2.0-3.0)
Candidate predictors	
How well can you walk?	9 (IQR 7-10)
How well can you use your hands?	9 (IQR 8-10)
Can you contain your urine well?	9 (IQR 7-10)
How good is your contact with members of your household?	10 (IQR 8-10)
How good is your memory?	8 (IQR 7-9)
Can you concentrate?	8 (IQR 7-9)
How do you feel?	8 (IQR 6-8)
Are you quickly fatigued?	7 (IQR 6-9)
Impairment of sensory tract	1 (IQR 1-2)
Impairment of pyramidal tract	1 (IQR 0-1)
Impairment of cerebellar tract	1 (IQR 0-2)
T2-weighted supratentorial lesion load (cm^3)	3.4 (IQR 0.8-11.3)
T2-weighted infratentorial lesion load (cm^3)	0.2 (IQR 0-0.5)
T2-weighted total lesion load (cm^3)	3.6 (IQR 1-11.4)
Number of lesions in spinal cord	2 (IQR 1-4)

SD = standard deviation; IQR = Inter-quartile range.

TABLE 6.3 Frequencies of unfavourable outcomes after 3 years (n=146).

	n	%
Inability to walk at least 500 m	37	25.3
Impaired dexterity	46	31.5
Cognitive impairments	44	30.1
Incontinence	24	16.4
Inability to use a car or public transportation	14	9.6
Social dysfunction	60	41.1
Reliance on a disability pension	77	52.7

istics of the patients, most of which comply with the expected pattern: more females than males, and approximately 80% with a relapse onset.

TABLE 6.3 shows the number of patients with an unfavourable outcome at the three-year follow-up. In the first three years walking ability decreased in 26 patients, 21 patients became incontinent, and 55 patients became (partially) reliant on a disability pension. In contrast to these deteriorations, remarkable improvement occurred in the cognitive impairment of 29 patients. The social functioning outcome showed important changes in both directions (21 patients improved and 22 deteriorated).

The final regression models, corrected for over-optimism by bootstrapping, are shown in TABLE 6.4. FIGURE 6.1 shows the discrimination and calibration curves. The outcomes for inability to walk at least 500 m, impaired dexterity and cognitive impairments show good calibration (the calibration curves follow approximately the 45° diagonal and the shrinkage factors [slope] approach 1). The calibration curves for the other outcomes show important miscalibration. Discriminitative ability is good for the models predicting inability to walk at least 500 metres (AUC=0.89 [0.83-0.95]), and incontinence (AUC=0.80 [0.71-0.90]), fair for the model predicting impaired dexterity (AUC=0.77 [0.69-0.86]), cognitive impairments (AUC=0.74), inability to use a car or public transportation (AUC=0.76 [0.65-0.83]) and reliance on a disability pension (AUC=0.72 [0.64-0.80]), and poor for the model predicting social dysfunction (AUC=0.67 [0.58-0.76]).

TABLE 6.4 Final regression models and their predictive ability.

Models and predictors (score range)		Predictive value β_{shrunk}	p	Model performance Slope	AUC
Inability to walk at least 500 m	How well can you walk (0-10)?	-0.57	0.00	0.93	0.89 (0.83-0.95)
	Impairment cerebellar tract (0-5)?	0.82	0.00		
	Number of lesions in spinal cord	0.17	0.05		
	Constant	2.31	0.05		
Impaired dexterity	How well can you use your hands (0-10)?	-0.19	0.16	0.85	0.77 (0.69-0.86)
	Impairment pyramidal tract (0-6)	0.30	0.31		
	Impairment cerebellar tract (0-5)?	0.53	0.03		
	Impairment sensory tract (0-6)?	0.31	0.17		
	T2-weighted infratentorial lesion load	1.14	0.00		
	Constant	-0.88	0.52		
Cognitive impairments	Age	0.03	0.12	0.88	0.74 (0.65-0.83)
	Gender	1.00	0.02		
	How well can you concentrate (0-10)?	-0.19	0.07		
	T2-weighted supratentorial lesion load	0.07	0.00		
	Constant	-1.82	0.12		
Incontinence	Can you contain your urine well (0-10)?	-0.45	0.00	0.97	0.80 (0.71-0.90)
	Number of lesions in spinal cord	0.10	0.25		
	Constant	1.54	0.06		
Inability to use a car or public transportation	How good is your memory (0-10)?	-0.27	0.06	0.71	0.76 (0.65-0.87)
	Impairment of pyramidal tract (0-6)	0.53	0.20		
	Impairment of cerebellar tract (0-5)	0.41	0.25		
	Number of lesions in spinal cord	0.17	0.09		
	Constant	-1.84	0.13		
Social dysfunction	How good is your contact with members of your household (0-10)?	-0.23	0.06	0.87	0.67 (0.58-0.76)
	Are you easily tired (0-10)?	0.19	0.01		
	T2-weighted total lesion load	0.02	0.27		
	Constant	0.35	0.79		
Reliance on a disability pension	How well can you concentrate (0-10)?	-0.27	0.01	0.84	0.72 (0.64-0.80)
	Impairment of pyramidal tract (0-6)	0.20	0.44		
	Impairment of cerebellar tract (0-5)	0.22	0.29		
	T2-weighted total lesion load	0.03	0.08		
	Constant	1.54	0.06		

Results for final models after internal validation by means of 250 bootstraps. $\beta_{shrunk} = \beta_{original}$ x slope. Slope = shrinkage factor obtained after bootstrapping; well calibrated models have a slope of 1. AUC (95%CI) = Area Under the receiver operation characteristic Curve with 95% Confidence Interval (95%CI); 0.50 indicates no discrimination beyond chance, >0.70 indicates sufficient discrimination.

Clinical prediction rules

Clinical prediction rules were constructed for the models predicting inability to walk at least 500 m, impaired dexterity and cognitive impairments (See APPENDIX).

Discussion

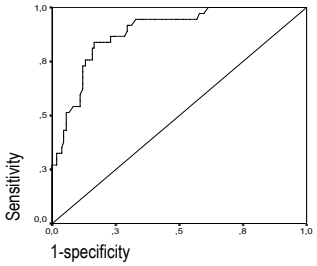
We have shown that it is feasible to make internally valid predictions for recently diagnosed MS patients with regard to outcomes on physical and cognitive functioning.

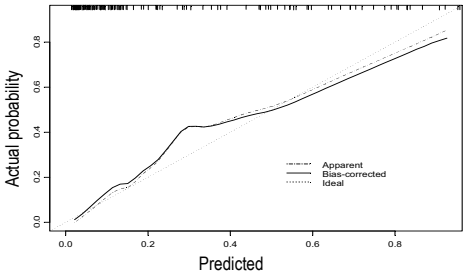
An important strength of our study is that the analysis was designed to optimize the internal validity.[28,29] Several attempts were made to minimize bias. Firstly, missing baseline data were imputed in order to optimize the quality of the data. Secondly, we used a limited set of clinically relevant candidate predictors that were only excluded when the p-value was greater than 0.50, or when the sign of the coefficient was opposite to what we expected. Finally, bootstrapping was used in order to correct for over-optimism of the regression coefficients and the model parameters (calibration: shrinkage factor, and discrimination: Area under the ROC curve).

A possible weakness of the study was the assessment of cognitive dysfunction. Twenty-nine patients showed cognitive improvements in the first three years, substantially more than the number of patients who improved on the other outcomes. In accordance with the design of our study,[14] cognitive data were collected annually, but it is possible that an interval of one year is not sufficiently long to rule out a practice effect. Another explanation might be that the definition of cognitive impairment that we applied does not correctly diagnose cognitively impaired patients. The cognitive screening test is based on five cognitive tests that each assess a different aspect of cognitive functioning, but in the literature there is no consensus on which cut-off point to use.[18,19,32-34] We used a sensitive cut-off point that classified patients as cognitively

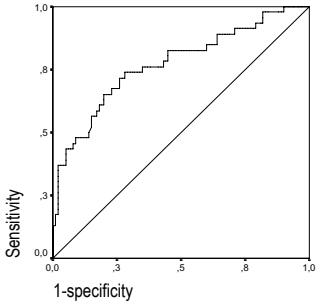
Discrimination	Calibration

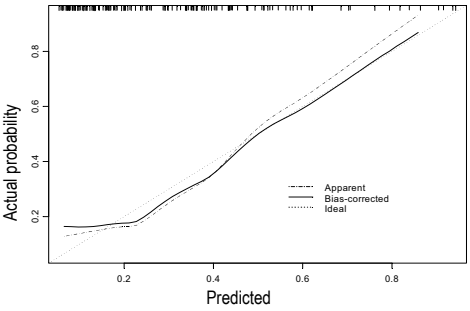
Inability to walk at least 500 metres



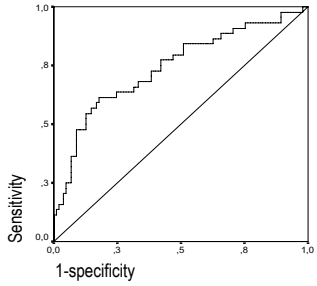


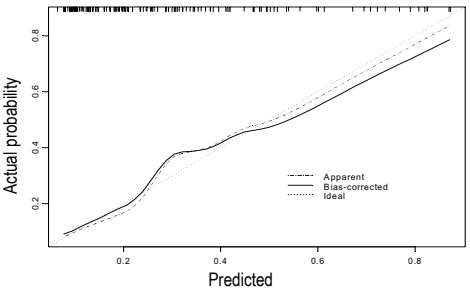
Impaired dexterity



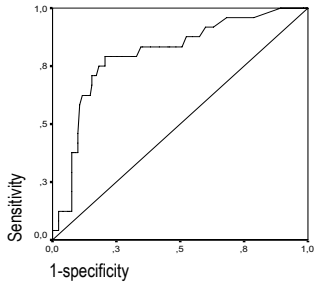


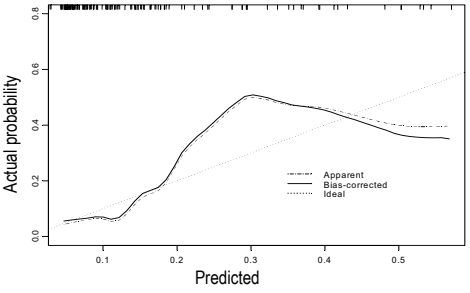
Cognitive impairments





Incontinence





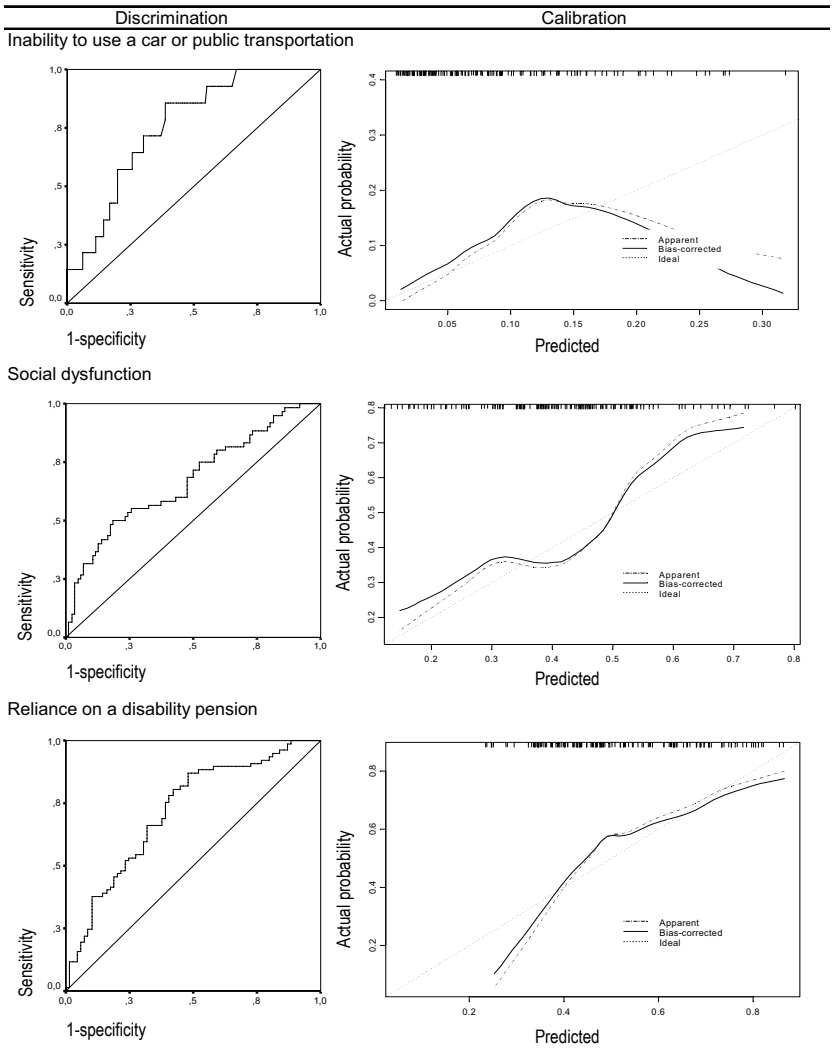


FIGURE 6.1 Discrimination and calibration curves for all outcomes. Discrimination: ROC curves that plot the sensitivity (true positive rate) vs. 1–specificity (false positive rate). The diagonal line indicates that there is no discrimination beyond chance, ideally the curve should approach the upper-left corner. Calibration: curves show the predicted vs. the observed probabilities. Ideally, the curve forms a 45° diagonal line.

impaired if one or more of their test-scores were lower than the mean – 1 x SD, compared to a Dutch reference population. Our strategy might therefore lead to a greater number of patients classified as cognitively impaired, whereas they actually perform within the norm, i.e. patients are classified as false-positive. Therefore, the observed improvements in cognitive functioning might just be changes that occur within normal ranges. Alternative cut-off points, such as two or more test-scores lower than the mean – 1 x SD, or one or more test-scores lower than the mean – 2 x SD, have also been applied in the literature. However, applying these criteria to our data still showed cognitive improvements for a substantial number of patients (data not shown). Therefore, the observed improvements in cognitive functioning are either caused by a practice effect or they are real improvements.

At baseline, i.e. a maximum of six months after the definite diagnosis was made, nine (6%) patients were receiving disease-modifying treatment. At the three-year follow-up, this rose to 44 (30%) with a mean treatment duration of 25 months. We did not include disease-modifying treatment at baseline in our models, because we assumed that confounding by indication could influence our findings. Patients with a more severe disease course are more likely to receive this treatment. The omission of disease-modifying treatment in the prediction models means that our models can be used independent of disease-modifying treatment. With regard to external validity, this means that our results can be generalized to populations in which approximately the same percentage of patient are receiving disease-modifying treatment.

Although our results look promising, application in clinical practice is not justified until they have been validated externally.[35-38] The analyses that we have presented should be repeated in a new cohort, which should be recruited in a different geographical area, at a different point in time, or, as is current in MS, assessed with different diagnostic criteria.[39] The regression coefficients and model parameters in these cohorts should be used to assess the

applicability of these models in clinical practice. When external validation has shown that the models perform well, and when the clinical usefulness of the clinical prediction rules has been established, they can be used with confidence in clinical practice to aid clinicians in making a prognosis. However, because the application of research findings in clinical practice is not self-evident, the clinical prediction rules should be actively implemented.[40,41]

From our results it can be seen that predictions of the outcomes that are based on performance measures, i.e. measures that require patients to actually perform a physical or cognitive test, are better than the predictions of outcomes based on self-reported health status. This implies that the more objective outcomes can be correctly predicted, but that self-reported outcomes are more difficult to predict. The reason for this might be that personal or social factors, which are not easy to measure as predictors, also have an effect on self-reported outcomes. In clinical practice, the clinical prediction rules could be used not only to improve treatment decisions regarding the initiation of disease-modifying treatment, but also to improve the timing of the (components of) rehabilitation treatment. Of equal importance is the possibility to improve the counselling of a patient. In conversations with the patient, the clinician should familiarize himself with the patient's personal and social situation, so that by combining this information with the information obtained from the clinical prediction rules, the clinician is able to formulate a patient-specific prognosis, which can then be discussed with the patient. The results of this discussion can be used to adjust the counselling of the patient, or can lead to the initiation of preventive measures or (rehabilitation) treatment.

Conclusion

In conclusion, during the first three years of MS it is possible to predict accurately inability to walk at least 500 m, impaired dexterity and cognitive impairments based on predictors that are deter-

mined shortly after the definite diagnosis has been made. The ability to predict physical and cognitive functioning might facilitate the counselling of patients and the planning of (rehabilitation) treatment. But first, adequate performance of the models in a new cohort must be validated externally.

References

1. Janssens AC, de Boer JB, van Doorn PA, van der Ploeg HM, van der Meche FG, Passchier J *et al.* Expectations of wheelchair-dependency in recently diagnosed patients with multiple sclerosis and their partners. Eur J Neurol 2003; 10:287-293.
2. Janssens AC, van Doorn PA, de Boer JB, van der Meche FG, Passchier J, Hintzen RQ. Perception of prognostic risk in patients with multiple sclerosis: the relationship with anxiety, depression, and disease-related distress. J Clin Epidemiol 2004; 57:180-186.
3. Runmarker B, Andersson C, Oden A, Andersen O. Prediction of outcome in multiple sclerosis based on multivariate models. J Neurol 1994; 241:597-604.
4. Runmarker B, Andersen O. Prognostic factors in a multiple sclerosis incidence cohort with twenty-five years of follow-up. Brain 1993; 116:117-134.
5. Weinshenker BG, Bass B, Rice GPA, Noseworthy J, Carriere W, Baskerville J *et al.* The natural history of multiple sclerosis: A geographically based study. 2. Predictive value of the early clinical course. Brain 1989; 112:1419-1428.
6. Weinshenker BG, Bass B, Rice GPA, Noseworthy J, Carriere W, Baskerville J *et al.* The natural history of multiple sclerosis: A geographically based study. I. Clinical course and disability. Brain 1989; 112:133-146.
7. Weinshenker BG, Rice GPA, Noseworthy JH, Carriere W, Baskerville J, Ebers GC. The natural history of multiple sclerosis: A geographically based study. 3. Multivariate analysis of predictive factors and models of outcome. Brain 1991; 114:1045-1056.
8. Amato MP, Ponziani G. A prospective study on the prognosis of multiple sclerosis. Neurol Sci 2000; 21:S831-S838.
9. Confavreux C, Vukusic S, Adeleine P. Early clinical predictors and progression of irreversible disability in multiple sclerosis: an amnesic process. Brain 2003; 126:770-782.
10. Pittock SJ, Mayr WT, McClelland RL, Jorgensen NW, Weigand SD, Noseworthy JH *et al.* Change in MS-related disability in a population-based cohort: a 10-year follow-up study. Neurology 2004; 62:51-59.
11. Harrison DA, Brady AR, Parry GJ, Carpenter JR, Rowan K. Recalibration of risk

prediction models in a large multicenter cohort of admissions to adult, general critical care units in the United Kingdom. Crit Care Med 2006; 34:1378-1388.

12. Matheny ME, Ohno-Machado L, Resnic FS. Discrimination and calibration of mortality risk prediction models in interventional cardiology. J Biomed Inform 2005; 38:367-375.
13. Poser CM, Paty DW, Scheinberg L, McDonald WI, Davis FA, Ebers GC *et al.* New diagnostic criteria for multiple sclerosis: guidelines for research protocols. Ann Neurol 1983; 13:227-231.
14. de Groot V, Beckerman H, Lankhorst GJ, Polman CH, Bouter LM. The initial course of daily functioning in multiple sclerosis: a three-year follow-up study. Mult Scler 2005; 11:713-718.
15. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology 1983; 33:1444-1452.
16. Uitdehaag BMJ, Ader HJ, Roosma TJ, de Groot V, Kalkers NF, Polman CH. Multiple sclerosis functional composite: impact of reference population and interpretation of changes. Mult Scler 2002; 8:366-371.
17. Rao SM, Leo GJ, Bernardin L, Unverzagt F. Cognitive dysfunction in multiple sclerosis. I. Frequency, patterns, and prediction. Neurology 1991; 41:685-691.
18. Boringa JB, Lazeron RH, Reuling IE, Adèr HJ, Pfennings L, Lindeboom J *et al.* The brief repeatable battery of neuropsychological tests: normative values allow application in multiple sclerosis clinical practice. Mult Scler 2001; 7:263-267.
19. Achiron A, Barak Y. Cognitive impairment in probable multiple sclerosis. J Neurol Neurosurg Psychiatry 2003; 74:443-446.
20. Granger CV, Cotter AC, Hamilton BB, Fiedler RC, Hens MM. Functional assessment scales: A study of persons with multiple sclerosis. Arch Phys Med Rehabil 1990; 71:870-875.
21. van Bennekom CAM, Jelles F, Lankhorst GJ, Bouter LM. The rehabilitation activities profile: A validation study of its use as a disability index with stroke patients. Arch Phys Med Rehabil 1995; 76:501-507.
22. Aaronson NK, Muller M, Cohen PD, Essink-Bot ML, Fekkes M, Sanderman R *et al.* Translation, validation, and norming of the Dutch language version of the SF-36 Health Survey in community and chronic disease populations. J Clin Epidemiol 1998; 51:1055-1068.
23. Laman H, Lankhorst GJ. Subjective weighting of disability: An approach to quality of life assessment in rehabilitation. Disabil Rehabil 1994; 16:198-204.
24. Lankhorst GJ, Jelles F, Smits RCF, Polman CH, Kuik DJ, Pfennings LEMA *et al.* Quality of life in multiple sclerosis: The disability and impact profile (DIP). J Neurol 1996; 243:469-474.

25. Kalkers NF, Bergers L, de Groot V, Lazeron RH, van Walderveen MA, Uitdehaag BM *et al.* Concurrent validity of the MS Functional Composite using MRI as a biological disease marker. Neurology 2001; 56:215-219.
26. Bot JC, Barkhof F, Polman CH, Nijeholt GJ, de Groot V, Bergers E *et al.* Spinal cord abnormalities in recently diagnosed MS patients: added value of spinal MRI examination. Neurology 2004; 62:226-233.
27. NORM: Multiple imputation of incomplete multivariate data under a normal model, version 2. Software for Windows 95/98/NT, available from http://www.stat.psu.edu/~jls/misoftwa.html. 1999.
28. Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol 1999; 52:935-942.
29. Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001; 54:774-781.
30. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996; 15:361-387.
31. Swets JA. Measuring the accuracy of diagnostic systems. Science 1988; 240:1285-1293.
32. Solari A, Mancuso L, Motta A, Mendozzi L, Serrati C. Comparison of two brief neuropsychological batteries in people with multiple sclerosis. Mult Scler 2002; 8:169-176.
33. Aupperle RL, Beatty WW, Shelton FN, Gontkovsky ST. Three screening batteries to detect cognitive impairment in multiple sclerosis. Mult Scler 2002; 8:382-389.
34. Dent A, Lincoln NB. Screening for memory problems in multiple sclerosis. Br J Clin Psychol 2000; 39:311-315.
35. Hukkelhoven CW, Rampen AJ, Maas AI, Farace E, Habbema JD, Marmarou A *et al.* Some prognostic models for traumatic brain injury were not valid. J Clin Epidemiol 2006; 59:132-143.
36. Bleeker SE, Moll HA, Steyerberg EW, Donders AR, Derksen-Lubsen G, Grobbee DE *et al.* External validation is necessary in prediction research: a clinical example. J Clin Epidemiol 2003; 56:826-832.
37. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000; 19:453-473.
38. Righini M, Bounameaux H. External validation and comparison of recently described prediction rules for suspected pulmonary embolism. Curr Opin Pulm Med 2004; 10:345-349.

39. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999; 130:515-524.
40. Grol R, Grimshaw J. From best evidence to best practice: effective implementation of change in patients' care. Lancet 2003; 362:1225-1230.
41. Bero LA, Grilli R, Grimshaw JM, Harvey E, Oxman AD, Thomson MA. Closing the gap between research and practice: an overview of systematic reviews of interventions to promote the implementation of research findings. The Cochrane Effective Practice and Organization of Care Review Group. BMJ 1998; 317:465-468.

APPENDIX Clinical Prediction Rules

		Score	Probability (%)
Inability to walk at least 500 m			
How well can you walk (0-10)?	x 3 =	> 22	0-20
Impairment of cerebellar tract (0-5)	x -5 =	22-16	20-40
Number of lesions in spinal cord	x -1 =	16-11	40-60
		11-5	60-80
	Σ Score = ☐	< 5	80-100
Impaired dexterity			
How well can you use your hands (0-10)?	x 1 =	> 3	0-20
Impairment of pyramidal tract (0-6)	x -2 =	3- -2	20-40
Impairment of cerebellar tract (0-5)	x -3 =	-2- -7	40-60
Impairment of sensory tract (0-6)	x -2 =	-7- -12	60-80
T2-weighted infratentorial lesion load (cm^3)	x -6 =	< -12	80-100
	Σ Score = ☐		
Cognitive impairments			
Age	x 1 =	< 14	0-20
Male	+ 29 =	14-47	20-40
How well can you concentrate (0-10)?	x -5 =	47-74	40-60
T2-weighted supratentorial lesion load (cm^3)	x 2 =	74-107	60-80
		> 107	80-100
	Σ Score = ☐		

Impairment of cerebellar tract
0 = normal
1 = abnormal signs without disability
2 = mild ataxia
3 = moderate truncal or limb ataxia
4 = severe ataxia
5 = unable to perform co-ordinated movements

Impairment of pyramidal tract
0 = normal
1 = abnormal signs without disability
2 = monoparesis grade 4
3 = monoparesis grade 2/3, or paraparesis/hemiparesis grade 3/4
4 = monoparesis grade 0/1, or paraparesis/hemiparesis grade 2, or tetraparesis grade 3/4
5 = paraparesis/hemiaresis grade 0/1, or tetraparesis grade 2
6 = tetraparesis grade 0/1

Impairment of sensory tract
0 = normal
1 = vibration/figure-writing mildly decreased 1-2 limbs
2 = vibration mildly decreased 3-4 limbs, or vibration moderately decreased 1-2 limbs, or touch/pain mildly decreased 1-2 limbs
3 = vibration lost 1-2 limbs, or touch/pain moderately decreased 1-2 limbs, or proprioception moderately decreased 3-4 limbs
4 = touch/pain moderately decreased 3-4 limbs, or touch/pain markedly decreased 1-2 limbs, or proprioception markedly decreased 3-4 limbs, or proprioception lost 1-2 limbs
5 = touch/pain moderately decreased below head, or sensation lost 1-2 limbs, or proprioception lost below the head
6 = sensation lost below the head

7

General discussion

Vincent de Groot

The aims of this thesis were to study the course of multiple sclerosis (MS) and its determinants, and the clinimetric properties of the outcome measures that can be used to study this disease course. Thus far, studies of MS have mainly focused on the disease course from a neurological point of view.[1-9] We studied it from a more functional perspective, and used the International Classification of Functioning (ICF) as model to study and explain the consequences of MS for the patient.[10] The ICF describes how patients live with their disease, and therefore investigates beyond mortality and disease. It is a classification of functioning that describes body functions and structures, and activities and participation. Because individual functioning and disability occur in a context, it also includes an assessment of environmental factors. The ICF can be used for research as well as for clinical practice purposes, and can also be used to classify outcome measures.

This thesis describes the clinimetric properties of several outcome measures that were used in our study, and the disease course in several domains of functioning. In this final CHAPTER we will critically discuss the main issues related to the study design, the clinimetric studies, and the studies of the disease course.

Methodological issues

Sampling

The majority of the patients were recruited from a tertiary referral centre, to which they had been referred to at their own request or at their physician's request. It is possible that relatively many distressed and severely affected patients were included in our study. If this method of selection had biased our results, this would mean that the domains of mobility and mental health would have been more severely affected. However, as we have shown in CHAPTER 4, the scores for the outcomes in these domains are normal, or almost normal, which makes selection bias less likely. Therefore, we are also confident that our results in the domain of social functioning are not severely biased.

Disease-modifying treatment

Ideally, in a longitudinal study the treatment of the patients should either be standardized or withheld. Of course, this cannot be justified in patients with a chronic progressive disease like MS, for which effective therapy was available and became available during the study.[11,12] The fact that patients were treated during the study may have influenced their outcome. Disease-modifying treatment reduces the rate of relapse by approximately 30%, but it has not yet convincingly been shown to reduce disability progression, and certainly not in such a short study period of three years.[11,12] At the baseline measurement, 6% of the patients, and at the last measurement 30% of the relapse onset patients were receiving disease-modifying treatment. In general, the patients who were receiving disease-modifying treatment were more severely affected (higher EDSS scores), which probably indicates that more severely affected patients are more likely to receive this treatment. The opposite interpretation, namely that patients receiving this treatment fare worse, is less likely, given the positive results of the systematic reviews.[11,12] If we assume a positive effect of disease-modifying treatment on disease progression, the progression we found in the domain of physical functioning for the relapse onset group may have been attenuated. However, given the equivocal results of the reviews, and in view of the fact that some of the patients received this treatment, we assume that this could have had no more than a weak effect on our results.

Missing data

In a longitudinal study like ours, which consists of 780 measurements (156 patients with five measurements each), it is inevitable that there is missing data. Because we made great efforts to minimize drop-out, only nine of the 156 patients were lost to follow-up. Apart from the missing measurements caused by this loss to follow-up, 15 other measurements were missing. This did not have an important effect on the results reported in CHAPTERS 4 and 5,

because we used longitudinal data-analysis techniques that are robust to small amounts of missing data.[13] In CHAPTER 6 we could only analyse the cases with a known 3-year outcome. In an effort to minimize any bias that may arise from analyses on a data-set with incomplete data on the determinants, we imputed missing MRI data in order to prevent further data loss.[14]

Outcomes

The selection of relevant outcomes was partly based on whether they had been used in previous research on MS patients. Therefore, the selected outcomes are a mixture of outcomes that were developed using the Health Related Quality of Life (HRQoL) model[15,16] and outcomes that were developed using the ICF as model.[10] Both models are widely used, but differ with regard to their theoretical framework. Although this may suggest that both models are mutually exclusive, it is to a certain extent possible to 'translate' HRQoL-based measures into the ICF model.[17-19] As outlined in the first paragraph of this CHAPTER, our analyses were based on the ICF. Therefore, we categorized the sub-scales of the Medical Outcome Study Short Form 36 (SF36)[20-23], a HRQoL measure, according to the ICF. The SF36 sub-scales Physical Functioning, Role Physical, Role Emotional and Social Functioning were classified as measures of activity and participation, and the SF36 sub-scales Vitality and Bodily Pain were classified as measures of basic functions. This categorization also explains why we used the SF36 sub-scales vitality and bodily pain as determinants of the social functioning outcome, as measured with the three social sub-scales of the SF36 (CHAPTER 5), although all sub-scales are part of the same HRQoL measure.

The use of the ICF model alos explains why we used the SF36 sub-scales Role Physical, Role Emotional, and Social Functioning to measure social functioning. Within the HRQoL model the sub-scale Role Physical belongs to the physical component of the SF36, and the sub-scales Role Emotional and Social Functioning belong to the mental component of the SF36.[21-24] The – ambiguous – general

question 'During the past four weeks, have you had any of the following problems *with your work or other regular daily activities* as a *result of your physical health?*' that proceeds the items of the SF36 sub-scale Role Physical, and the question 'During the past four weeks, have you had any of the following problems *with your work or other regular daily activities* as a *result of any emotional problems?*' that proceeds the items of the SF36 sub-scale Role Emotional, cause confusion as to what is really measured with these items. The main goal of these items is to measure limitations in social functioning. Because the limitations in social functioning (*i.e.* your work or other regular daily activities) result from either physical health or emotional problems, it may seem that the two sub-scales belong to a different dimension of the SF36. However, according to the ICF model, both questions measure limitations in social functioning, although the underlying causes of these limitations may be different. In our view, the addition of 'as a result of …' causes unnecessary ambiguity in these questions.

A phenomenon that occurs in longitudinally measured patient-assessed outcomes is response shift.[25,26] However, research on the clinical relevance of this phenomenon is not yet conclusive.[27] In our study, response shift may play a role. Although some outcomes did not change, it is still possible that in reality there was a change in these domains. Respondents might have adapted to their situation and rate their current health situation against the background of newly adapted standards. Such a response shift can involve changing internal standards of measurement (scale recalibration), values (*i.e.* the importance of components constituting the target construct) or redefining the target construct itself.[25,26] Several methods can be used to assess response shift.[28] Although most of them require a specific design that is tailored to assess response shift, there are some methods that can be integrated in longitudinal studies. Statistical methods look for differences in factor structure of the questionnaire at different points in time, because these differences in factor structure

may indicate response shift. These methods can be used when very large sample sizes are available. The Then-test requires a design in which the questionnaire is completed three times: the first time at baseline, the second time at follow-up evaluating the current situation, and the third time also at follow-up, but then re-evaluating the situation at baseline. Differences between the answers to the first and the third questionnaire are indicative of response shift, since the latter was completed against the background of newly adapted standards. In our study, each outcome measure was only measured once at each follow-up measurement, so we did not have a Then-test in our study.

The increased attention that is being paid to response shift has also led to fundamental questions about the measurement of Quality of Life (QoL) and the proposal of a new psychometric model.[29,30] The observations that QoL ratings are quite unpredictable, and dependent on numerous (unknown) factors have given rise to these questions.[30] In our opinion, there are two possible approaches to this problem: 1) consider QoL as a highly variable, but valid and directly measurable construct, or 2) consider QoL as a more stable latent construct that can not be measured directly, because it is highly dependent on individual appraisal processes.

Schwartz and Rapkin[29] describe how future research should progress when based on the second approach. They distinguish three types of constructs to measure: performance-based, perception-based and evaluation-based. In performance-based measurements objective data are collected, in perception-based measurements the patient is asked about objective performance, and in evaluation-based measurements the patient is asked about the difficulties experienced with a specific task. They suggest that for QoL measurement one is interested in the individual evaluation, because this evaluation contains some form of appraisal. Furthermore, the new psychometric model should focus strongly on the appraisal process that underlies the scoring. Ultimately, the QoL rating should be corrected for the individual appraisal process

in order to achieve QoL ratings that are comparable across studies and patients. Although we strongly advise that the proposed levels of measurement should be distinguished, and we agree with Schwarz and Rapkin that it is important to improve our understanding of the appraisal processes, we believe that this approach does still not accurately measure QoL.

Why not adopt the first approach and consider QoL as a highly variable, but valid and directly measurable construct? A closer look at the items of the SF36 reveals that numerous constructs are being measured, with the underlying assumption that they all measure aspects of QoL. However, the SF36 does not contain any items that directly assess QoL. When a patient indicates that he/she finds it difficult to perform a specific task (*e.g.* I have a lot of difficulty climbing the stairs), we do not have sufficient information to know how the patient's QoL will be affected. It may be that part of the problem in the assessment of QoL is due to insufficient content validity of the current QoL measures. In our opinion, QoL might as well be a construct that stands on its own, and is not necessarily coupled to the performance of a task, the perception of the performance of a task, or even the evaluation of the performance of a task, or impairments such as pain and fatigue. Instead of developing a new psychometric model, it may be possible to develop a single construct QoL measure that consists of items that directly assess QoL. This QoL measure could then be evaluated, based on the traditional psychometric measurement model, with the ultimate goal to develop a unidimensional single construct QoL measure based on the Item Response Theory. Subsequently, this single construct QoL measure could be used in studies investigating which factors contribute to QoL in individual patients. The consequence of this approach may be that performance-based or perception-based measures turn out to be more suitable for evaluation of changes in clinical trials than such a single construct QoL measure.

Determinants

We assessed a wide variety of determinants that may be associated with our outcomes in the relevant domains. For most of the constructs that we intended to measure it was fairly easy to find suitable, validated and reliable outcome measures. However, a suitable outcome measure for the construct comorbidity[31] could not easily be found. Therefore, we carried out a systematic review to evaluate the clinimetric properties of the methods that are available for the assessment of comorbidity.[32] CHAPTER 2 describes the results of this review, and shows that the clinimetric properties of four of the methods that we identified, are sufficient to make them applicable for research. The choice between these four depends on the outcome one is interested in and the target population. For our study, we chose the Cumulative Illness Rating Scale (CIRS)[33], because it can be used independent of the outcome, and can easily be applied in different populations. Furthermore, it is relatively easy to administer in a patient interview, because it is organized around the major body systems, includes a severity rating per body system, and does not depend on specific diagnoses, which makes chart review superfluous.

Main findings

Clinimetrics

The most important longitudinal studies of the course of MS used the Expanded Disability Status Scale (EDSS)[34] as outcome.[1,3,4,6,7,9,35] The EDSS was developed by Kurtzke in 1967 to assess the severity of the disease.[34] He managed to construct a sensible outcome measure, because despite all the criticism of the EDSS that emerged from the recent advancements in clinimetrics,[36-38] it still is a scale that is very familiar for clinicians and researchers. It facilitates communication between clinicians, and is used to characterize the study population in research papers. So, even nowadays, it provides an estimate (however crude) of the severity of MS.

Of course, the current criticism should be acknowledged, and attempts should be made to overcome the shortcomings of the EDSS, either by improving the EDSS, which has not yet been attempted, or by developing new outcome measures. Since 1998, when we started our longitudinal study, several new outcome measures that aim to quantify the impact or severity of MS have been developed.[39-50]Some of these newly developed outcomes may turn out to be welcome additions or improvements to the already available MS-specific outcomes. The MS Functional Composite (MSFC) has been developed to increase power in clinical trials.[39] A task-force reviewed a number of trials that used other outcomes than the EDSS, and selected three continuous outcomes, related to mobility, dexterity and cognition, to assess the impact of MS on the patient. The three outcomes were transformed, using Z-scores, and subsequently summed to form a total score. However, the usefulness of this measure is still under debate,[48,51-53] because of the difficulty in interpreting change scores and the problem concerning practice effects of the Paced Auditory Serial Addition Test, a component of the MSFC. Three Randomized Controlled Trials (RCT) have used the MSFC as primary outcome.[40-42] The Guy's Neurological Disability Scale (GNDS)[54] assesses the disability that arises from important MS-related symptoms, such as fatigue, mobility, and bowel and bladder function. It has been found to be acceptable in clinimetric evaluations, and two RCTs [43,44] have used the GNDS as primary outcome. Two other scales, the MS Impact Scale (MSIS)[45] and the MS Spasticity Scale (MSSS)[46] have been developed according to sound clinimetric methods. In other evaluation studies the clinimetric properties of the MSIS look promising.[47-50] No studies have yet reported on the MSSS, and these two measures have not yet been used in any RCTs.

In CHAPTER 3 we assessed the usefulness of evaluative outcome measures for MS.[53] Our approach differs from those described above. Instead of trying to assess the overall impact of MS on functioning, *i.e.* constructing an MS-specific outcome, we used

the ICF model to categorize our outcomes. This means that we classified our outcomes according to the activity and participation domains of the ICF, and because the ICF distinguishes basic functions and activities, we categorized the outcome measures as much as possible according to the ICF constructs.

What does this mean for the selection of outcomes? Imagine the following three patients. Patient one is limited in his social functioning, because he has a severe paraparesis. Patient two is equally limited in his social functioning, because he is very fatigued. Patient three is equally limited in his social functioning, because he is very depressed. According to the ICF we need four outcome measures to be able to distinguish these patients: an outcome at participation level to measure social functioning, and three outcomes at the level of basic functions to assess paresis, fatigue and depression. Each outcome should measure its construct as pure as possible, which means that the questions should be unambiguous. Outcomes that measure the impact of MS may be based on questions like: 'to what extent does MS limit your social functioning'. This causes problems for patient 3, because he is not sure whether his social functioning is limited due to MS or due to depression. He may indicate that MS does not limit his social functioning, which may be right if he has been depressed for a very long time (longer than he has had MS). But what if his depression occurred after the diagnosis of MS? We recommend that conditional questions should be avoided when assessing a particular outcome, and that each relevant construct should be measured as pure as possible.

One may argue that MS-specific outcomes are more tailored to the disease and its consequences, and may therefore be better in detecting changes (in RCTs). As we have shown in CHAPTER 3, this hypothesis should be tested in studies that compare several outcomes with each other.[53] Our study had already started before the most recently developed outcomes measures became available, so we were not able to test them. However, our results indicate that

some ICF domain-specific outcomes, such as the Medical Outcome Study Short Form 36 sub-scale physical functioning[20] and the Rehabilitation Activities Profile subscale occupation,[55,56] perform better than the disease-specific outcomes, such as the EDSS[34] and the MSFC[39], especially in the first three years after diagnosis. Our results indicate that the choice of an outcome measure should be based on the domains that will probably be affected in the population under study, and that it is not mandatory to select an MS-specific outcome. Especially in the early stage of MS, when neurological symptoms are relatively mild, one will not find much with the MSFC, even though it is a continuous outcome that has the potential to be responsive. Furthermore, using single domain outcomes facilitates the interpretation of the results.

Another important point that arises from CHAPTER 3 is that none of the outcome measures is capable of detecting changes in individual patients, at least in the early stages of the disease. This is a very important finding, because the current trend in individual patient care is to rely increasingly more on outcome measures.[57] Our results clearly indicate that the requirements of outcomes for use in individual patient care are higher than the requirements of outcomes for use in research. This is in contrast with the general opinions of clinicians. Therefore, our results are not only relevant for early MS, but can also be used to design studies addressing the use of outcome measures in individuals with more severe MS or other diseases.

Functional prognosis

Our study provides a unique comprehensive description of the course of MS in several relevant domains of functioning.[58] In addition to providing a different way of investigating the course of MS, it also reintroduces powerful statistical techniques to analyze longitudinal data in MS research.[59] These techniques make it possible to draw sound conclusions from relatively small studies on the clinical course of MS. Results show that 'the' disease course does not

exist, but that one should consider the consequences of MS on several different domains of functioning. In CHAPTER 4 we have shown that, although neurological impairments are relatively mild and mental health is relatively mildly affected, early in the course of MS social functioning is seriously affected.[58] Although these findings are not really new to clinicians with experience in the treatment of patients with MS, this is the first longitudinal study that provides clear data in support of this clinical observation. The extent to which social functioning is affected is greater than we anticipated, and the relatively limited effect on mental functioning shortly after the diagnoses was made, was also unexpected. We assumed that mental health would be markedly affected shortly after the diagnosis was made. After this we expected it to gradually improve, although we did not expect it to return to (almost) normal levels.

In CHAPTER 5 we investigated the possible determinants related to social dysfunctioning. From an epidemiological point of view it makes sense to specify a theoretical model of the expected relationships before the analysis is undertaken. Preferably, one should choose a limited number of determinants of interest and their potential confounders in order to determine the strength of the relationship between these central determinants and the outcome with the least possible bias. We used a theoretical model, and categorized our determinants in four clusters that were crudely based on the ICF model: patient and disease characteristics, basic functions, psychosocial characteristics and basic abilities. However, because the literature implied that many determinants may be associated with social dysfunctioning, we wanted to find out which of these determinants are the most strongly or consistently associated with social dysfunctioning. Therefore, we decided not to limit our selection of determinants, but explicitly decided to compare all potentially relevant determinants. Because comparing a large number of determinants introduces the risk of bias, we used several outcome measures for social functioning to investigate whether the selection of determinants was robust across these

outcomes, and we performed a sensitivity analysis with a more liberal p-value to investigate whether the choice of determinants was influenced by adherence to a strict p-value. Determinants that were selected for several outcomes, and that also had the strongest associations with the outcomes were considered to be the most important determinants. In our opinion, this study has provided the basis for improved understanding of social dysfunctioning in the early stages of MS, and has identified the three most important theoretical concepts that are related to social functioning: disease activity, vitality or fatigue, and social support.

In CHAPTER 6 we focused on the predictability of the course of MS in the first three years. The results indicate that it seems feasible to accurately predict the individual course of MS for the outcomes of mobility, dexterity and cognition. The methodology that is used for this kind of research assesses the accuracy of the predictions, and not the strength of the predictors.[60,61] This analytical approach is totally different from the approach described in the previous paragraph, where the focus is on the strength of the determinants. To our knowledge there are no studies of MS that specifically focus on the accuracy of the predictions themselves. In the literature the two approaches are often not clearly distinguished from each other, often because of ambiguity in the research question that is addressed. Clarity in the literature will be promoted if research questions based on increasing our understanding of underlying mechanisms are investigated with techniques that address the strength of the relationships, preferably in longitudinal studies, and research questions about predictions are investigated in longitudinal studies with measures assessing the accuracy of the predictions of the outcomes.

Future research

Clinimetrics

The results reported in CHAPTER 3 indicated that some measures are very suitable for research projects.[53] However, they also show

that the clinimetric properties are not sufficient to allow the application of these outcome measures in individual patient care. Future studies should try to improve the clinimetric properties of outcome measures so that they can be applied in individual patient care, and to improve the responsiveness of the outcome measures. A potential technique is Rash analysis, a technique that is based on the Item Response Theory. It assesses whether scales are unidimensional, and transforms the scores into an interval scale. This has the potential to improve the responsiveness of these measures, and concurrently may improve the properties that currently limit their applicability in individual patient care. Until now, we have only investigated, and confirmed, that some of the outcome measures we used fit the Rasch model.[62,63] The next step will be to investigate whether the responsiveness of Rasch transformed scales is superior to that of non-Rasch-transformed scores.

We studied the responsiveness in patients in the early stages of MS. With continuing follow-up we will have the opportunity to study the responsiveness in more severely disabled MS patients. These analyses may show different results, because the relative relevance of affected domains will change with increasing disability.

Another clinimetric issue concerns the use of outcome measurement in clinical practice, especially related to the use of the ICF in patients with MS. Since the publication of the ICF, core-sets that consist of ICF items that are particularly relevant for a diagnostic group have been developed for several (neurological) diagnostic groups.[64-68] When these core-sets are used to measure changes over time in individual patients, either to monitor progression or to measure treatment effects, the clinimetric demands regarding the reliability and responsiveness of these measures are high.[57] Furthermore, when the aim is to apply these measures in clinical practice there is also a requisite that these measure are easy to apply, and not too costly in time or resources. In our opinion, this trend of developing ICF core-sets should be accompanied by the selection (if outcomes are already available) or development of

short, easy to apply, highly reliable, highly responsive, single construct measures. Rasch analyses may be very helpful in the development of these measures. Ultimately, this should result in a set of short, high quality, single construct measures that cover all relevant domains of the ICF. In clinical practice, the clinician can select the outcome measures that are relevant for the clinical decision-making process for each individual patient. In our view, the choice should be predominantly based on the problems encountered in contacts with the patient. Mandatory outcomes measurements in clinical practice should serve a clear goal that is agreed upon by all relevant stake-holders.

Functional prognosis

The 3-year follow-up of the cohort described in this thesis is relatively short, and this is reflected to some extent in the relatively mild deterioration in the domains of mobility and mental health.[58] Because in this progressive neurological disease the patient's disability increases with the increasing length of follow-up, extension of the follow-up of this cohort will almost certainly show further deterioration in these domains. The data on the determinants that have been and will be collected, can then be analyzed with longitudinal analysis techniques in order to identify the most relevant determinants for deterioration in individual patients in these domains. This is valuable, because it informs us about the relationship between changes in determinants and changes in the outcomes in individual patients (within subject effect), which is much stronger evidence of a relationship than associations found in cross-sectional studies that can only inform us about between subject effects.

The determinants that are identified with these longitudinal analyses should then be used to develop randomized clinical trials that manipulate these determinants. In CHAPTER 5 we identified four important determinants in three areas in which RCTs can be carried out. The area of disease activity is covered by the determi-

nants T2-weighted lesion load and the number of self-reported exacerbations. Current disease-modifying treatment and RCTs in this area have a strong focus on manipulating these determinants. Two other relevant determinants are perceived social support and vitality. Multidisciplinary rehabilitation treatment to optimize perceived social support and vitality should be developed and tested in separate RCTs. A lack of perceived social support may be effectively treated by altering coping mechanisms. The negative effects of reduced vitality on functioning may be treated effectively with endurance training,[69-71] or energy conservation programmes[72,73]. In other patient groups, such as patients with chronic fatigue syndrome[74] and cancer patients with fatigue[75], positive effects of cognitive behavioral therapy have been reported. Instead of teaching patients to optimize the balance between capacity and workload on the basis of their fatigue, which reinforces their focus on their fatigue, patients are taught to gradually improve their daily (social) functioning on the basis of a time-contingent schedule, while at the same time changing their cognition of their fatigue, *i.e.* they learn that their life does not have to be ruled by their fatigue. Although cognitive behavioral therapy has not yet been investigated with regard to fatigue in MS, it has been investigated, and found to be relatively effective, as treatment for depression and acceptation problems in patients with MS.[76] In our opinion, rehabilitation interventions should no longer be studied as complete all-encompassing packages, because it has already been shown that rehabilitation is effective for MS.[77-79] Instead, new efforts should focus on establishing the effectiveness of the components of these treatments,[80-82] such as those described above. It will be a challenge to conduct an RCT to study the effects of specific components of a rehabilitation intervention in a population that is characterized by multiple co-occurring problems.

In CHAPTER 6 we have shown that it seems feasible to construct prediction rules over a period of three years for the outcomes of mobility, dexterity and cognition. Before widespread

implementation can be aimed for, these prediction rules should first be validated in a new cohort.[83-86] Furthermore, we should establish whether the predictive ability is maintained when the period over which the predictions are made increases. Another question relates to the use of clinical prediction rules in clinical practice. Beattie and Nelson[87] concluded that clinical prediction rules can be of great value to assist clinical decision-making, but should not be used indiscriminately, they are not a replacement for clinical judgment, and should complement rather than supplant clinical opinion and intuition. Reilly and Evans[88] made a useful distinction between assistive prediction rules and directive decision rules. Assistive prediction rules only provide the probability of a certain outcome without attaching a specific decision to this probability, whereas directive decision rules provide the user with a clear decision as outcome. They also showed that there are decision rules that are superior to the decisions made by clinicians, which may imply that these decision rules must be adhered to. However, respecting the autonomy of the clinicians, they argue that the use of directive decision rules should be discretionary and not mandatory. Furthermore, even if high methodological standards have been met, they suggest that the effect of the clinical decision rules on the outcome should be evaluated by performing a formal impact analysis afterwards.[88] The conclusions of Beattie and Nelson[87] apply to assistive prediction rules, but are too weak for directive decision rules with a proven capacity to outperform clinicians. Although we have shown that prediction in MS seems feasible, at this point in time we should be modest in our recommendations to apply our assistive prediction rules in clinical practice. Therefore, there should be no widespread implementation of our assistive prediction rules before external validation has been performed. The implementation should then only take place with the conclusions of Beattie and Nelson[87] in mind and, thus, accompanied by instructions on how to apply the rules in practice.

CHAPTER 7

Clinical implications

Clinimetrics

The use of outcome measures in individual patient care is widely advocated. In our opinion, there are five reasons for this advice: 1) to improve the diagnostic process, 2) to improve prognostication, 3) to monitor disease progression, 4) to evaluate treatment effects, and 5) to generate management information. The clinimetric demands of the outcome measures that are selected depend on the reasons for measurement. Outcome measures that are used to improve the diagnostic process should be able to discriminate cases from non-cases, and, therefore, be judged on the basis of characteristics such as sensitivity, specificity, and positive and negative predictive value.[89] For the evaluation of these outcome measures reference data from healthy control subjects are necessary. Outcome measures that are intended to improve prognostication should generally be fairly easy to obtain in regular individual patient care, and should have been tested for their ability to improve the prediction of future outcomes. In practice, this means that two large cohorts are needed, one for the development of the prediction rule and one for its validation, before evidence-based application in clinical practice can be recommended. Outcome measures that are used to measure disease progression or to evaluate treatment effects in clinical practice should be able to reliably identify clinically relevant changes in individual patients, which means that the demands with regard to reliability and responsiveness are high, in comparison with outcome measures that are used for research purposes. Outcome measures to support management decisions should be selected on the basis of relevance for the involved stake-holders. Their clinimetric requirements should probably be equal to those needed for research.

In this thesis we did not assess outcome measures with regard to their diagnostic value, but we did try to improve prognostication. The results of CHAPTER 6 indicate that it may be possible to

predict mobility, dexterity and cognitive disorders three years after the diagnosis has been made, based on predictors obtained at baseline. At this moment, we cannot yet provide clinicians with any evidence-based advice regarding the application of our prediction rules, because a formal validation should first be performed. However, because many patients think that they are at risk for a worse prognosis (*i.e.* being wheelchair-bound within a few years),[90] even though large-scale group studies[1-9] and our results in CHAPTER 6 indicate otherwise, we suggest that clinicians use this information to inform their patients about the difference between the patient's risk perception and the risk as derived from cohort studies. Although caution is necessary, we suggest that clinicians implicitly, without naming percentages, incorporate this – short term – prognostic information about mobility, dexterity and cognition when holding a conversation with the patient about the consequences of MS on personal (social) functioning in the near future, with the aim to adjust the patient's risk perception of a worse outcome. Stating that MS is unpredictable, probably reinforces a patient's risk perception of a worse outcome, and may cause unnecessary distress.

In CHAPTER 3 we tried to identify outcome measures that can be used to evaluate changes. Unfortunately, none of the outcome measures we tested were able to reliably evaluate changes at individual level. Clinicians who apply outcome measures to monitor disease progression or to evaluate changes in the early stages of MS should be aware of these limitations. Because mobility is an important domain in which progression takes place, and the SF36 sub-scale mobility and the 10-meter timed walk test are not suitable to evaluate changes in individual patients, it makes sense to use an outcome measure that has more potential to evaluate individual changes, such as the 6-minute walk test[91]. For the evaluation of individualized (rehabilitation) treatment goals Goal Attainment Scaling may be used.[92]

Functional prognosis

Approximately 40% of the MS patients have problems with social functioning. This means that clinicians who treat these patients in the early stages of the disease should actively address social functioning. Furthermore, we have shown that disease activity (exacerbations and lesion loads), vitality, and perceived social support are the main determinants of this dysfunctioning. The implications of these findings for clinical practice may be that disease activity should be treated, vitality should be improved, for example by promoting physical activity in order to maintain or obtain a healthy life-style (at least 5 days per week 30 minutes of moderate physical activity), and patients should be informed about the role of perceived social support and how to cope with and mobilize social support.

References

1. Runmarker B, Andersen O. Prognostic factors in a multiple sclerosis incidence cohort with twenty-five years of follow-up. Brain 1993; 116:117-134.
2. Runmarker B, Andersson C, Oden A, Andersen O. Prediction of outcome in multiple sclerosis based on multivariate models. J Neurol 1994; 241:597-604.
3. Weinshenker BG, Bass B, Rice GPA, Noseworthy J, Carriere W, Baskerville J *et al.* The natural history of multiple sclerosis: A geographically based study. I. Clinical course and disability. Brain 1989; 112:133-146.
4. Weinshenker BG, Bass B, Rice GPA, Noseworthy J, Carriere W, Baskerville J *et al.* The natural history of multiple sclerosis: A geographically based study. 2. Predictive value of the early clinical course. Brain 1989; 112:1419-1428.
5. Weinshenker BG, Rice GPA, Noseworthy JH, Carriere W, Baskerville J, Ebers GC. The natural history of multiple sclerosis: A geographically based study. 4. Applications to planning and interpretation of clinical therapeutic trials. Brain 1991; 114:1057-1067.
6. Amato MP, Ponziani G. A prospective study on the prognosis of multiple sclerosis. Neurol Sci 2000; 21:S831-S838.
7. Confavreux C, Vukusic S, Adeleine P. Early clinical predictors and progression of irreversible disability in multiple sclerosis: an amnesic process. Brain 2003; 126:770-782.
8. Pittock SJ, Mayr WT, McClelland RL, Jorgensen NW, Weigand SD, Noseworthy JH

et al. Disability profile of MS did not change over 10 years in a population-based prevalence cohort. Neurology 2004; 62:601-606.

9. Weinshenker BG, Rice GPA, Noseworthy JH, Carriere W, Baskerville J, Ebers GC. The natural history of multiple sclerosis: A geographically based study. 3. Multivariate analysis of predictive factors and models of outcome. Brain 1991; 114:1045-1056.
10. WHO. International Classification of Functioning, Disability and Health: ICF. Geneva: WHO, 2001.
11. Rice GP, Incorvaia B, Munari L, Ebers G, Polman C, D'Amico R *et al.* Interferon in relapsing-remitting multiple sclerosis. Cochrane Database Syst Rev 2001;CD002002.
12. Filippini G, Munari L, Incorvaia B, Ebers GC, Polman C, D'Amico R *et al.* Interferons in relapsing remitting multiple sclerosis: a systematic review. Lancet 2003; 361:545-552.
13. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics 1986; 42:121-130.
14. NORM: Multiple imputation of incomplete multivariate data under a normal model, version 2. Software for Windows 95/98/NT, available from http://www.stat.psu.edu/~jls/misoftwa.html. 1999.
15. Romney DM, Evans DR. Toward a general model of health-related quality of life. Qual Life Res 1996; 5:235-241.
16. Anderson RT, Aaronson NK, Wilkin D. Critical review of the international assessments of health- related quality of life. Qual Life Res 1993; 2:369-395.
17. Cieza A, Stucki G. Content comparison of health-related quality of life (HRQOL) instruments based on the international classification of functioning, disability and health (ICF). Qual Life Res 2005; 14:1225-1237.
18. Borchers M, Cieza A, Sigl T, Kollerits B, Kostanjsek N, Stucki G. Content comparison of osteoporosis-targeted health status measures in relation to the International Classification of Functioning, Disability and Health (ICF). Clin Rheumatol 2005; 24:139-144.
19. Stucki G, Cieza A. The International Classification of Functioning, Disability and Health (ICF) Core Sets for rheumatoid arthritis: a way to specify functioning. Ann Rheum Dis 2004; 63 Suppl 2:ii40-ii45.
20. Ware JE, Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992; 30:473-483.
21. Freeman JA, Hobart JC, Langdon DW, Thompson AJ. Clinical appropriateness: a key factor in outcome measure selection: the 36 item short form health survey in multiple sclerosis. J Neurol Neurosurg Psychiatry 2000; 68:150-156.
22. Anderson C, Laubscher S, Burns R. Validation of the Short Form 36 (SF-36)

health survey questionnaire among stroke patients. Stroke 1996; 27:1812-1816.

23. Aaronson NK, Muller M, Cohen PD, Essink-Bot ML, Fekkes M, Sanderman R *et al.* Translation, validation, and norming of the Dutch language version of the SF-36 Health Survey in community and chronic disease populations. J Clin Epidemiol 1998; 51:1055-1068.
24. Ware JE Jr., Kosinski M, Gandek B, Aaronson NK, Apolone G, Bech P *et al.* The factor structure of the SF-36 Health Survey in 10 countries: results from the IQOLA Project. International Quality of Life Assessment. J Clin Epidemiol 1998; 51:1159-1165.
25. Schwartz CE, Sprangers MA. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. Soc Sci Med 1999; 48:1531-1548.
26. Sprangers MA, Schwartz CE. Integrating response shift into health-related quality of life research: a theoretical model. Soc Sci Med 1999; 48:1507-1515.
27. Schwartz CE, Bode R, Repucci N, Becker J, Sprangers MA, Fayers PM. The clinical significance of adaptation to changing health: a meta-analysis of response shift. Qual Life Res 2006; 15:1533-1550.
28. Visser MR, Oort FJ, Sprangers MA. Methods to detect response shift in quality of life data: a convergent validity study. Qual Life Res 2005; 14:629-639.
29. Schwartz CE, Rapkin BD. Reconsidering the psychometrics of quality of life assessment in light of response shift and appraisal. Health Qual Life Outcomes 2004; 2:16.
30. Rapkin BD, Schwartz CE. Toward a theoretical model of quality-of-life appraisal: Implications of findings from studies of response shift. Health Qual Life Outcomes 2004; 2:14.
31. Feinstein AR. The pre-therapeutic classification of co-morbidity in chronic disease. J Chronic Dis 1970; 23:455-468.
32. de Groot V, Beckerman H, Lankhorst GJ, Bouter LM. How to measure comorbidity. a critical review of available methods. J Clin Epidemiol 2003; 56:221-229.
33. Linn BS, Linn MW, Gurel L. Cumulative illness rating scale. J Am Geriatr Soc 1968; 16:622-626.
34. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: an expanded disability status scale (EDSS). Neurology 1983; 33:1444-1452.
35. Pittock SJ, Mayr WT, McClelland RL, Jorgensen NW, Weigand SD, Noseworthy JH *et al.* Change in MS-related disability in a population-based cohort: a 10-year follow-up study. Neurology 2004; 62:51-59.
36. Noseworthy JH. Clinical scoring methods for multiple sclerosis. Ann Neurol 1994; 36:S80-S85.

37. Sharrack B, Hughes RA, Soudain S, Dunn G. The psychometric properties of clinical rating scales used in multiple sclerosis. Brain 1999; 122:141-159.
38. Hobart J, Freeman J, Thompson A. Kurtzke scales revisited: the application of psychometric methods to clinical intuition. Brain 2000; 123:1027-1040.
39. Cutter GR, Baier ML, Rudick RA, Cookfair DL, Fischer JS, Petkau J *et al.* Development of a multiple sclerosis functional composite as a clinical trial outcome measure. Brain 1999; 122:871-882.
40. Romberg A, Virtanen A, Ruutiainen J. Long-term exercise improves functional impairment but not quality of life in multiple sclerosis. J Neurol 2005; 252:839-845.
41. Schwid SR, Petrie MD, Murray R, Leitch J, Bowen J, Alquist A *et al.* A randomized controlled study of the acute and chronic effects of cooling therapy for MS. Neurology 2003; 60:1955-1960.
42. Cohen JA, Cutter GR, Fischer JS, Goodman AD, Heidenreich FR, Kooijmans MF *et al.* Benefit of interferon beta-1a on MSFC progression in secondary progressive MS. Neurology 2002; 59:679-687.
43. Craig J, Young CA, Ennis M, Baker G, Boggild M. A randomised controlled trial comparing rehabilitation against standard therapy in multiple sclerosis patients receiving intravenous steroid treatment. J Neurol Neurosurg Psychiatry 2003; 74:1225-1230.
44. Wade DT, Young CA, Chaudhuri KR, Davidson DL. A randomised placebo controlled exploratory study of vitamin B-12, lofepramine, and L-phenylalanine (the 'Cari Loder regime') in the treatment of multiple sclerosis. J Neurol Neurosurg Psychiatry 2002; 73:246-249.
45. Hobart J, Lamping D, Fitzpatrick R, Riazi A, Thompson A. The Multiple Sclerosis Impact Scale (MSIS-29): a new patient-based outcome measure. Brain 2001; 124:962-973.
46. Hobart JC, Riazi A, Thompson AJ, Styles IM, Ingram W, Vickery PJ *et al.* Getting the measure of spasticity in multiple sclerosis: the Multiple Sclerosis Spasticity Scale (MSSS-88). Brain 2006; 129:224-234.
47. Hobart JC, Riazi A, Lamping DL, Fitzpatrick R, Thompson AJ. How responsive is the Multiple Sclerosis Impact Scale (MSIS-29)? A comparison with some other self report scales. J Neurol Neurosurg Psychiatry 2005; 76:1539-1543.
48. Hoogervorst EL, Zwemmer JN, Jelles B, Polman CH, Uitdehaag BMJ. Multiple Sclerosis Impact Scale (MSIS-29): relation to established measures of impairment and disability. Mult Scler 2004; 10:569-574.
49. McGuigan C, Hutchinson M. The multiple sclerosis impact scale (MSIS-29) is a reliable and sensitive measure. J Neurol Neurosurg Psychiatry 2004; 75:266-269.
50. van der Linden FA, Kragt JJ, Klein M, van der Ploeg HM, Polman CH, Uitdehaag

BM. Psychometric evaluation of the multiple sclerosis impact scale (MSIS-29) for proxy use. J Neurol Neurosurg Psychiatry 2005; 76:1677-1681.

51. Solari A, Radice D, Manneschi L, Motti L, Montanari E. The multiple sclerosis functional composite: different practice effects in the three test components. J Neurol Sci 2005; 228:71-74.
52. Kragt JJ, van der Linden FA, Nielsen JM, Uitdehaag BM, Polman CH. Clinical impact of 20% worsening on Timed 25-foot Walk and 9-hole Peg Test in multiple sclerosis. Mult Scler 2006; 12:594-598.
53. de Groot V, Beckerman H, Uitdehaag BM, de Vet HC, Lankhorst GJ, Polman CH *et al.* The usefulness of evaluative outcome measures in patients with multiple sclerosis. Brain 2006; 129:2648-2659.
54. Sharrack B, Hughes RA. The Guy's Neurological Disability Scale (GNDS): a new disability measure for multiple sclerosis. Mult Scler 1999; 5:223-233.
55. van Bennekom CAM, Jelles F, Lankhorst GJ. Rehabilitation activity profile – A brief research report on an application of the ICIDH in rehabilitation medicine. Eur J Phys Med Rehabil 1993; 3:125-128.
56. van Bennekom CAM, Jelles F, Lankhorst GJ. Rehabilitation Activities Profile: The ICIDH as a framework for a problem-oriented assessment method in rehabilitation medicine. Disab Rehabil 1995; 17:169-175.
57. Dekker J, Dallmeijer AJ, Lankhorst GJ. Clinimetrics in rehabilitation medicine: current issues in developing and applying measurement instruments 1. J Rehabil Med 2005; 37:193-201.
58. de Groot V, Beckerman H, Lankhorst GJ, Polman CH, Bouter LM. The initial course of daily functioning in multiple sclerosis: a three-year follow-up study. Mult Scler 2005; 11:713-718.
59. D'yachkova Y, Petkau J, White R. Longitudinal analyses for magnetic resonance imaging outcomes in multiple sclerosis clinical trials. J Biopharm Stat 1997; 7:501-531.
60. Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol 1999; 52:935-942.
61. Steyerberg EW, Harrell FE, Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001; 54:774-781.
62. Dallmeijer AJ, Dekker J, Roorda LD, Knol DL, van Baalen B, de Groot V *et al.* Differential item functioning of the Functional Independence Measure in higher performing neurological patients. J Rehabil Med 2005; 37:346-352.
63. Dallmeijer AJ, Dekker J, Knol DL, Kalmijn S, Schepers VP, de Groot V *et al.* Dimensional structure of the SF-36 in neurological patients. J Clin Epidemiol 2006; 59:541-543.

64. Biering-Sorensen F, Scheuringer M, Baumberger M, Charlifue SW, Post MW, Montero F *et al.* Developing core sets for persons with spinal cord injuries based on the International Classification of Functioning, Disability and Health as a way to specify functioning. Spinal Cord 2006; 44:541-546.
65. Coenen M, Cieza A, Stamm TA, Amann E, Kollerits B, Stucki G. Validation of the International Classification of Functioning, Disability and Health (ICF) Core Set for rheumatoid arthritis from the patient perspective using focus groups. Arthritis Res Ther 2006; 8:R84.
66. Stier-Jarmer M, Grill E, Ewert T, Bartholomeyczik S, Finger M, Mokrusch T *et al.* ICF Core Set for patients with neurological conditions in early post-acute rehabilitation facilities. Disabil Rehabil 2005; 27:389-395.
67. Ewert T, Grill E, Bartholomeyczik S, Finger M, Mokrusch T, Kostanjsek N *et al.* ICF Core Set for patients with neurological conditions in the acute hospital. Disabil Rehabil 2005; 27:367-373.
68. Geyh S, Cieza A, Schouten J, Dickson H, Frommelt P, Omar Z *et al.* ICF Core Sets for stroke. J Rehabil Med 2004;135-141.
69. Petajan JH, Gappmaier E, White AT, Spencer MK, Mino L, Hicks RW. Impact of aerobic training on fitness and quality of life in multiple sclerosis. Ann Neurol 1996; 39:432-441.
70. Oken BS, Kishiyama S, Zajdel D, Bourdette D, Carlsen J, Haas M *et al.* Randomized controlled trial of yoga and exercise in multiple sclerosis. Neurology 2004; 62:2058-2064.
71. Surakka J, Romberg A, Ruutiainen J, Aunola S, Virtanen A, Karppi SL *et al.* Effects of aerobic and strength exercise on motor fatigue in men and women with multiple sclerosis: a randomized controlled trial. Clin Rehabil 2004; 18:737-746.
72. Finlayson M. Pilot study of an energy conservation education program delivered by telephone conference call to people with multiple sclerosis. NeuroRehabilitation 2005; 20:267-277.
73. Mathiowetz VG, Finlayson ML, Matuska KM, Chen HY, Luo P. Randomized controlled trial of an energy conservation course for persons with multiple sclerosis. Mult Scler 2005; 11:592-601.
74. Prins JB, van der Meer JW, Bleijenberg G. Chronic fatigue syndrome. Lancet 2006; 367:346-355.
75. Gielissen MF, Verhagen S, Witjes F, Bleijenberg G. Effects of cognitive behavior therapy in severely fatigued disease-free cancer patients compared with patients waiting for cognitive behavior therapy: a randomized controlled trial. J Clin Oncol 2006; 24:4882-4887.
76. Thomas PW, Thomas S, Hillier C, Galvin K, Baker R. Psychological interventions for multiple sclerosis. Cochrane Database Syst Rev 2006;CD004431.
77. Thomson TD, Taylor WJ. Evidence for inpatient rehabilitation as an effective

intervention. Hosp Med 2005; 66:200-204.
78. Kesselring J. Neurorehabilitation in multiple sclerosis—what is the evidence-base? J Neurol 2004; 251 Suppl 4:IV25-IV29.
79. Thompson AJ. The effectiveness of neurological rehabilitation in multiple sclerosis. J Rehabil Res Dev 2000; 37:455-461.
80. Steultjens EM, Dekker J, Bouter LM, Cardol M, van de Nes JC, van den Ende CH. Occupational therapy for multiple sclerosis. Cochrane Database Syst Rev 2003;CD003608.
81. Thompson AJ. Neurorehabilitation in multiple sclerosis: foundations, facts and fiction. Curr Opin Neurol 2005; 18:267-271.
82. Rietberg MB, Brooks D, Uitdehaag BM, Kwakkel G. Exercise therapy for multiple sclerosis. Cochrane Database Syst Rev 2005;CD003980.
83. Hukkelhoven CW, Rampen AJ, Maas AI, Farace E, Habbema JD, Marmarou A *et al.* Some prognostic models for traumatic brain injury were not valid. J Clin Epidemiol 2006; 59:132-143.
84. Bleeker SE, Moll HA, Steyerberg EW, Donders AR, Derksen-Lubsen G, Grobbee DE *et al.* External validation is necessary in prediction research: a clinical example. J Clin Epidemiol 2003; 56:826-832.
85. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000; 19:453-473.
86. Righini M, Bounameaux H. External validation and comparison of recently described prediction rules for suspected pulmonary embolism. Curr Opin Pulm Med 2004; 10:345-349.
87. Beattie P, Nelson R. Clinical prediction rules: what are they and what do they tell us? Aust J Physiother 2006; 52:157-163.
88. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med 2006; 144:201-209.
89. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM *et al.* Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ 2003; 326:41-44.
90. Janssens AC, van Doorn PA, de Boer JB, van der Meche FG, Passchier J, Hintzen RQ. Perception of prognostic risk in patients with multiple sclerosis: the relationship with anxiety, depression, and disease-related distress. J Clin Epidemiol 2004; 57:180-186.
91. Perera S, Mody SH, Woodman RC, Studenski SA. Meaningful change and responsiveness in common physical performance measures in older adults. J Am Geriatr Soc 2006; 54:743-749.
92. Hurn J, Kneebone I, Cropley M. Goal setting as an outcome measure: A systematic review. Clin Rehabil 2006; 20:756-772.

Summary

This thesis describes the results of the FuPro MS study (Functional Prognosis in Multiple Sclerosis) that was financially supported by the Netherlands Organization for Scientific Research (NWO 940-33-009), and that was carried out in close cooperation with the Functional Prognostication and disability (FuPro) study group. In the period 1998-2000 all consecutive potentially eligible patients visiting the outpatient clinics of neurology departments of the VU University Medical Center, the Academic Medical Center Amsterdam, the Sint Lucas Andreas Hospital Amsterdam, the OLVG Hospital Amsterdam, and the Erasmus Medical Center Rotterdam were invited to participate. We recruited a cohort of 156 recently (less than six months previously) diagnosed patients, aged 16 to 55 years. The cohort was prospectively monitored for five measurements in three years. Diagnosis was determined according to the Poser-criteria for definite Multiple Sclerosis (MS). Patients with other neurological disorders, systemic diseased or malignant neoplastic diseases were excluded. The patients were visited at home in order to minimise drop-out, and four well-trained raters were responsible for the scoring. The aims of this thesis were to study the disease course of MS and its determinants, and the clinimetric properties of the outcome measures that can be used to study this disease course.

CHAPTER 2 describes a systematic review of available methods to measure comorbidity and to assess their validity and reliability. We searched Medline and Embase, with the keywords comorbidity and multi-morbidity, to identify articles in which a method to

measure comorbidity was described. We also checked the references of these articles for other relevant articles. With a standardized checklist we extracted the relevant data from these articles. Subsequently, we assessed the content, concurrent, predictive and construct validity, and the reliability of the methods to measure comorbidity.

We identified thirteen different methods to measure comorbidity: one disease-count and 12 indexes. Data on content and predictive validity were available for all measures, while data on construct validity were available for nine methods, data on concurrent validity and inter-rater reliability for eight methods, and data on intra-rater reliability for three methods.

It was concluded that the Charlson Index, the Cumulative Ilness Rating Scale (CIRS), the Index of Coexisting Disease (ICED) and the Kaplan Index were valid and reliable methods to measure comorbidity that can be used in clinical research. The Charlson Index is the most extensively studied comorbidity index for predicting mortality. The CIRS addresses all relevant body systems without using specific diagnoses. The ICED has a two-dimensional structure, measuring disease severity and disability, which can be useful when mortality and disability are the outcomes of interest. The Kaplan Index was specifically developed for use in diabetes research. For the other indexes, insufficient data on the clinimetric properties were available.

In CHAPTER 3 we aimed to select the most useful evaluative outcome measures for early multiple sclerosis (MS). All 156 recently diagnosed MS patients were included in a three-year follow-up study, and assessed on 23 outcome measures in the domains of disease-specific outcomes, physical functioning, mental health, social functioning, and general health. A Global Rating Scale (GRS) and the Expanded Disability Status Scale (EDSS) were used as external criteria to determine the Minimally Important Change (MIC) for each outcome measure. Subsequently, we determined whether the outcome measures could detect their MIC reliably.

From these, per domain the outcome measure that was found to be most sensitive to changes (responsive) was identified.

We found that at group level, eleven outcomes of the domains of physical functioning, mental health, social functioning and general health could reliably detect the MIC. Of these eleven, the most responsive measures per domain were the Medical Outcome Study 36 Short Form sub-scale physical functioning (SF36pf), the Disability and Impact Profile (DIP) sub-scale psychological, the Rehabilitation Activities Profile sub-scale occupation (RAPocc), and the DIP sub-scale mental health, respectively. Overall, the most responsive measures were the SF36pf and the RAPocc. However, in individual patients, none of the measures could reliably detect the MIC.

We concluded that in the early stages of MS the most useful evaluative outcome measures for research are the SF36pf (physical functioning) and the RAPocc (social functioning).

In CHAPTER 4 we studied the initial course of daily functioning in multiple sclerosis. All longitudinally gathered data of the whole cohort of 156 recently diagnosed patients in the domains neurological deficits, physical functioning, mental health, social functioning and general health were used in the analysis. We made an a priori distinction between a relapse onset group (n=128) and a non-relapse onset group (n=28).

We showed that at baseline, neurological deficits are relatively minor for most patients, 26.3% have aberrant physical functioning scores, 38.5% have aberrant social functioning scores, 9.0% have aberrant mental health scores and 25.0% have aberrant general health scores. In the subsequent three years the neurological deficits and physical functioning deteriorated significantly. This deterioration was more pronounced and clinically relevant in the non-relapse onset group. Mental health showed a significant, but not clinically relevant deterioration in this period. Social functioning and general health showed non-significant changes in this period of three years.

We concluded that in the initial stage of multiple sclerosis, when neurological deficits are relatively minor and mental health is relatively unaffected, patients in both groups experience limitations in daily functioning. Patients in the non-relapse onset group have progressive neurological symptoms that are accompanied by progressive limitations in physical functioning, but not by progressive limitations in the other domains.

CHAPTER 5 shows the results of a detailed analysis of the determinants of social functioning. The aim was to identify the strongest determinants of social functioning in the first three years after the diagnosis multiple sclerosis (MS) has been made. We used all longitudinally collected data in the cohort of 156 recently diagnosed MS patients. Social functioning was measured using the three social subscales of the Medical Outcome Study Short Form 36 (SF36) and a composite outcome based on these subscales. Forty-three, longitudinally collected, determinants were divided into the following clusters: 12 patient and disease characteristics, ten psychosocial characteristics, 18 basic functions, and three basic activities. The results were analyzed with generalized estimating equations (GEE), using the following steps: 1) a stepwise backwards selection procedure for all clusters per outcome, 2) an overall stepwise backwards selection procedure for each outcome using the significant variables identified in step one, 3) examine whether the associations are based on within subject changes, *i.e.* is a change in the determinant for a patient associated with a change in the outcome for that same patient, and 4) a sensitivity analysis.

In the first step of the analysis 17 determinants were selected in any of the four models. In step 2 we showed that vitality, the number of self-reported exacerbations and the perceived amount of social support were associated with social functioning in three or four of the models. In the next step we showed that almost all associations are based on within subject changes. In the final step the T2-weighted supratentorial lesion load was additionally selected.

Summarizing, vitality, the perceived amount of social support, and disease activity, *i.e.* the number of self-reported exacerbations and the T2-weighted supratentorial lesion load, are determinants of social functioning. These results suggest that, in the early stages of MS, it might be beneficial to improve vitality, *e.g.* by promoting an active life-style, to optimize the perceived amount of social support, *e.g.* by counselling, and to suppress disease activity, *e.g.* by prescribing interferon.

CHAPTER 6 investigated whether it is possible to accurately predict functioning after three years in patients with recently diagnosed MS. For this study we used the data of 146 recently diagnosed MS patients. For the other ten patients data of the outcome measurement at three years were missing. At baseline, predictors were obtained from history-taking, neurological examination and Magnetic Resonance Imaging. The outcomes of interest after three years were: inability to walk at least 500 metres, impaired dexterity, cognitive impairments, incontinence, inability to drive a car or use public transportation, social dysfunction, and reliance on a disability pension. We constructed clinical prediction rules for the models that were well calibrated, *i.e.* that showed sufficient agreement between predicted and observed outcomes, based on visual inspection of calibration curves, and that showed sufficient discrimination, *i.e.* the Area Under the receiver operation characteristic Curve (AUC) > 0.70, after internal bootstrap validation.

The results indicated that the models for inability to walk at least 500 metres, impaired dexterity and cognitive impairments were well calibrated. Discrimination was sufficient for all seven models, except for the one predicting social dysfunction (AUC = 0.67).

So we have shown that inability to walk at least 500 metres, impaired dexterity and cognitive impairments can be predicted with predictors that are obtained shortly after the definite diagnosis has been made. The ability to predict physical and cognitive functioning might facilitate the counselling of patients with MS and the planning of their (rehabilitation) treatment.

In CHAPTER 7 we discuss the methodological issues related to our study, the main findings, the recommendations for future research and the implications for clinical practice. We argue that potential sources of bias have a minor effect in our carefully documented cohort with few missing data. The main findings regarding the use of outcome measures in research and clinical practice are placed in the contexts of the ICF, generic versus disease specific outcome measures, and the use of outcome measures in groups versus individual patients. In the sections that discuss the main findings regarding functional prognosis we describe the difference between the analysis of longitudinal data with the intention to understand the relationships, *i.e.* try to ascertain whether a relationship may be causal, and the analysis of longitudinal data with the intention to predict certain outcomes. Recommendations for future research relate to responsiveness of outcome measures in individual patients, functional prognosis in this cohort with extended follow-up, Randomized Controlled Trials to develop treatments for the determinants that are related to social functioning, and validation studies of our prediction rules. In the section on clinical implications we give suggestions on how to deal with prognostic information in individual patients, discuss the limitations regarding the use of outcome measures to evaluate treatment effects in individual patient care, and urge clinicians to discuss the social consequences of MS with the patient in the early stage of MS.

Samenvatting

In dit proefschrift worden de resultaten van het FuPro MS onderzoek (Functionele Prognose bij patiënten met Multipele Sclerose) beschreven. Dit onderzoek werd gefinancierd door de Nederlandse organisatie voor Wetenschappelijk Onderzoek (NWO 940-33-009), en in nauwe samenwerking met de onderzoeksgroep functionele prognose bij chronische neurologische aandoeningen uitgevoerd. In de periode van 1998 tot 2000 werden alle achtereenvolgende patiënten die de poliklinieken neurologie van het VU medisch centrum, het Academisch Medisch Centrum, het Sint Lucas-Andreas ziekenhuis, het OLVG ziekenhuis en het Erasmus medisch centrum in Rotterdam bezochten, gevraagd deel te nemen aan het onderzoek. Honderdzesenvijftig recent (minder dan zes maanden geleden) gediagnosticeerde patiënten werden in het onderzoek ingesloten, en prospectief gevolgd gedurende drie jaar. In deze periode werden vijf metingen verricht. De diagnose werd gesteld aan de hand van de Poser-criteria voor zekere multipele sclerose (MS). Patiënten met andere neurologische aandoeningen, systemische ziekten of oncologische aandoeningen werden uitgesloten van deelname. De patiënten werden thuis bezocht om de uitval zoveel mogelijk te beperken. Vier goed getrainde onderzoekers verrichtten de metingen. In dit proefschrift worden het ziekteverloop van MS, de determinanten van dit verloop, en de klinimetrische eigenschappen van de meetinstrumenten die gebruikt worden om dit ziekteverloop te meten, bestudeerd.

HOOFDSTUK 2 bevat een systematische review van beschikbare methoden om comorbiditeit te meten. Van iedere meetmeth-

ode wordt de validiteit en betrouwbaarheid bepaald. We doorzochten Medline en Embase met de trefwoorden comorbiditeit en multimorbiditeit om artikelen te identificeren waarin een meetmethode voor comorbiditeit wordt beschreven. De referenties van de gevonden artikelen werden bestudeerd om andere artikelen te identificeren. Met behulp van een standaard checklist werden de relevante gegevens uit de artikelen verzameld. Daarna werd van iedere methode de inhoud, concurrente, predictieve en construct validiteit, en de betrouwbaarheid bepaald.

Dertien methoden om comorbiditeit te meten werden geïdentificeerd: één methode waar eenvoudigweg het aantal comorbide aandoeningen werd geteld, en twaalf indexen. Voor alle methoden waren gegevens over inhoud en predictieve validiteit beschikbaar. Voor negen methoden waren gegevens over construct validiteit beschikbaar. Voor acht methoden waren gegevens over concurrente validiteit en interbeoordelaarsbetrouwbaarheid beschikbaar. Voor drie methoden waren gegevens over intrabeoordelaarsbetrouwbaarheid beschikbaar.

We concludeerden dat de Charlson index, the Cumulative Illness Rating Scale (CIRS), the Index of Coexistent Disease (ICED) en de Kaplan index valide en betrouwbare methoden zijn om comorbiditeit te meten in klinische onderzoeksprojecten. De Charlson Index is de meest uitgebreid bestudeerde comorbiditeitsindex om mortaliteit te voorspellen. De CIRS omvat alle orgaansystemen zonder daarbij gebruik te maken van specifieke diagnosen. De ICED heeft een tweedimensionele structuur, waarmee ziekte-ernst en beperkingen gemeten worden. Dit kan goed bruikbaar zijn wanneer mortaliteit en beperkingen de uitkomsten zijn waar het onderzoek zich op richt. De Kaplan Index is speciaal ontwikkeld voor gebruik in onderzoek bij diabetes mellitus. Voor de overige indexen waren er onvoldoende gegevens beschikbaar om de klinimetrische eigenschappen te kunnen beoordelen.

In HOOFDSTUK 3 was het doel om de meest bruikbare meetinstrumenten te selecteren voor het meten van veranderin-

gen in de vroege fase van MS. Alle honderdzesenvijftig recent gediagnosticeerde MS patiënten werden geïncludeerd in een drie jaar durend onderzoek, en gemeten met behulp van drieëntwintig meetinstrumenten uit de domeinen ziekte-specifieke uitkomsten, fysiek functioneren, mentale gezondheid, sociaal functioneren, en algemene gezondheid. Een Global Rating Scale (globale beoordelingsschaal, GRS) en de Expanded Disability Status Scale (EDSS) werden gebruikt als externe criteria om de Minimal Important Change (minimaal relevante verandering, MIC) voor ieder meetinstrument te bepalen. Vervolgens bepaalden we of het meetinstrument zijn MIC betrouwbaar kan detecteren. Van de meetinstrumenten die in staat waren hun MIC betrouwbaar te bepalen werd vervolgens per domein het meetinstrument dat het gevoeligst is om veranderingen te meten (responsief) geïdentificeerd.

Voor de bestudering van groepen patiënten bleken elf uitkomstmaten uit de domeinen fysiek functioneren, mentale gezondheid, sociaal functioneren en algemene gezondheid in staat om hun MIC betrouwbaar te detecteren. Van deze elf waren de Medical Outcome Study 36 Short Form subschaal physical functioning (SF36pf), de Disability and Impact Profile (DIP) subschaal psychological, het Revalidatie Activiteiten Profiel subschaal dagbesteding (RAPdb) en de DIP subschaal mental health de meest responsieve uitkomstmaten per domein. Over het geheel genomen waren de SF36pf en de RAPdb de meest responsieve uitkomstmaten. Echter, geen van de meetinstrumenten was in staat om in een individuele patiënt de MIC betrouwbaar te bepalen.

We concludeerden dat in de vroege fase van MS de SF36pf (fysiek functioneren) en de RAPdb (sociaal functioneren) de meest bruikbare meetinstrumenten zijn om veranderingen te meten in onderzoek.

In HOOFDSTUK 4 bestudeerden we het initiële verloop van het dagelijks functioneren bij patiënten met MS. Longitudinaal verzamelde gegevens van het gehele cohort van honderdzesenvijftig

recent gediagnosticeerde patiënten uit de domeinen neurologische stoornissen, fysiek functioneren, mentale gezondheid, sociaal functioneren en algemene gezondheid werden gebruikt in de analyse. We maakten een *apriori* onderscheid tussen een groep waarbij de ziekte met een schub begon (relapse onset, n=128) en een groep waarbij dit niet het geval was (non-relapse onset, n=28).

Tijdens de eerste meting (binnen een half jaar na het stellen van de diagnose) zijn de neurologische stoornissen relatief mild voor de meeste patiënten, en ervaart 26,3% problemen met fysiek functioneren, 38,5% met sociaal functioneren, 9,0% met hun mentale gezondheid en 25,0% met hun algemene gezondheid. In de daaropvolgende drie jaar verslechteren de neurologische stoornissen en het fysiek functioneren in beide groepen; in de non-relapse onset groep is deze verslechtering meer uitgesproken en klinisch relevant. De mentale gezondheid liet een significante, maar niet klinisch relevante verslechtering zien voor beide groepen. Het sociaal functioneren en de algemene gezondheid veranderden niet significant in beide groepen.

Wij concludeerden dat in de initiële fase van MS, als de neurologische stoornissen nog relatief mild zijn en de mentale gezondheid nauwelijks is aangedaan, patiënten in beide groepen beperkingen in het dagelijks functioneren ervaren. Patiënten in de non-relapse onset groep hebben echter meer uitgesproken en progressieve neurologische verschijnselen die gepaard gaan met progressieve beperkingen in het fysiek functioneren, maar niet met progressieve beperkingen op andere domeinen.

In HOOFDSTUK 5 laten we de resultaten zien van een gedetailleerde analyse van de determinanten van sociaal disfunctioneren. Het doel was om de sterkste determinanten van sociaal functioneren in de eerste drie jaar na het stellen van de diagnose MS te identificeren. Alle longitudinaal verzamelde gegevens van het cohort van 156 patiënten werden gebruikt. Sociaal functioneren werd gemeten met behulp van de drie sociale subschalen van de Medical Outcome Study 36 Short Form (SF36) en een

samengestelde uitkomstmaat die gebaseerd is op deze drie subschalen. Drieënveertig, grotendeels longitudinaal gemeten, determinanten werden verdeeld over de volgende vier clusters: twaalf patiënt- en ziektekarakteristieken, tien psychosociale karakteristieken, achttien basisfuncties, en drie basisvaardigheden. De resultaten werden stapsgewijs geanalyseerd met behulp van Generalized Estimating Equations (GEE). Eerst werden door middel van een *backwards* selectieprocedure de belangrijkste determinanten per cluster en per uitkomstmaat geselecteerd. In de tweede stap werd per uitkomstmaat bepaald – opnieuw met behulp van een *backwards* selectieprocedure – welke van deze determinanten, die afkomstig zijn uit verschillende clusters, het meest belangrijk zijn. In de derde stap onderzochten we of veranderingen in een determinant samenhangen met veranderingen in de uitkomst voor diezelfde patiënt (intra-individuele veranderingen). Tot slot verrichtten we een sensitiviteitsanalyse.

In de eerste stap werden in totaal zeventien determinanten geselecteerd. In de tweede stap vonden we dat vitaliteit, het aantal door de patiënt zelf gerapporteerde exacerbaties en de ervaren hoeveelheid sociale steun geassocieerd waren met sociaal functioneren in drie van de vier modellen. In de volgende stap vonden we dat bijna alle gevonden associaties berustten op intra-individuele veranderingen. In de laatste stap werd ook het T2-gewogen supratentoriële laesievolume geselecteerd.

Samenvattend kunnen we stellen dat vitaliteit, de ervaren hoeveelheid sociale steun, en ziekteactiviteit, dat wil zeggen het aantal door de patiënt zelf gerapporteerde exacerbaties, en het T2-gewogen supratentoriële laesievolume, determinanten zijn van een verminderd sociaal functioneren. Onze resultaten suggereren dat, in de vroege fase van MS, het nuttig kan zijn om de vitaliteit te verbeteren, bijvoorbeeld door een actievere levensstijl te adviseren, de ervaren hoeveelheid sociale steun te optimaliseren, bijvoorbeeld door counseling, en ziekteactiviteit te onderdrukken, bijvoorbeeld door het gebruik van interferon.

In HOOFDSTUK 6 onderzochten we of het mogelijk is om het functioneren drie jaar na het stellen van de diagnose MS accuraat te voorspellen. Voor dit onderzoek gebruikten we de gegevens van 146 recent gediagnosticeerde MS patiënten. Voor de andere 10 patiënten ontbraken de gegevens over de situatie drie jaar na het stellen van de diagnose. Op baseline verzamelden we gegevens over de predictoren door middel van anamnese, neurologisch onderzoek en Magnetic Resonance Imaging (MRI). De zeven uitkomsten op drie jaar waren: onvermogen om tenminste 500 meter te lopen, ge- stoorde handfunctie, cognitieve stoornissen, incontinentie, onvermogen om auto te rijden of gebruik te maken van het openbaar vervoer, sociaal disfunctioneren, en geheel of gedeeltelijk gebruik maken van de WAO. We construeerden klinische predictieregels voor de modellen die, na interne *bootstrap* validatie, goed gekalibreerd waren (goede overeenstemming tussen voorspelde en waargenomen uitkomst), en goed konden discrimineren (gebied onder de *receiver operating characteristic curve* (AUC) groter dan 0,70).

De resultaten lieten zien dat de modellen voor het onvermogen om tenminste 500 meter te kunnen lopen, gestoorde handfunctie en cognitieve stoornissen goed gekalibreerd zijn. Discriminatie was voldoende voor alle modellen met uitzondering van het model dat sociaal disfunctioneren voorspelt (AUC = 0,67).

We toonden aan dat het onvermogen om tenminste 500 meter te kunnen lopen, gestoorde handfunctie en cognitieve stoornissen voorspeld kunnen worden met behulp van predictoren die rond het stellen van de definitieve diagnose MS zijn verkregen. De mogelijkheid om het fysiek en cognitief functioneren te voorspellen zou gebruikt kunnen worden om de counseling van patiënten met MS en het plannen van (revalidatie) behandeling te verbeteren.

In HOOFDSTUK 7 bediscussiëren we de methodologische kwesties gerelateerd aan onze studie, de belangrijkste bevindingen, de aanbevelingen voor toekomstig onderzoek en de impli-

caties voor de klinische praktijk. We beargumenteren dat de mogelijke vormen van bias een gering effect hebben in ons zorgvuldig gedocumenteerd cohort met weinig missende gegevens. De belangrijkste bevindingen met betrekking tot het gebruik van uitkomstmaten in onderzoek en klinische praktijk worden besproken in relatie tot de *International Classification of Functioning* (ICF), het generiek versus ziektespecifiek meten, en het gebruik van uitkomstmaten in groepen versus individuele patiënten. In de secties over de functionele prognose maken we duidelijk onderscheid tussen de analyse van longitudinale gegevens gericht op het begrijpen van onderliggende relaties en de analyse van longitudinale gegevens met het doel uitkomsten te voorspellen. Aanbevelingen voor toekomstig onderzoek gaan over de responsiviteit van uitkomstmaten in individuele patiënten, lange termijn functionele prognose van ons cohort, gerandomiseerde klinische trials gericht op het ontwikkelen van behandelingen voor de determinanten die geassocieerd zijn met sociaal disfunctioneren, en validatiestudies voor onze predictieregels. In de sectie over de klinische implicaties geven we suggesties hoe om te gaan met prognostische informatie in individuele patiënten, bespreken we de beperkingen van het gebruik van uitkomstmaten voor het evalueren van behandeleffecten in de individuele patiëntenzorg, en moedigen we clinici aan om al in de vroege fase van MS met de patiënt in gesprek te gaan over de ervaren sociale gevolgen van MS.

Dankwoord

Weet je het zeker? Het gaat wel over een periode van acht jaar. Dat was de strekking van de vraag die ik kreeg voorgelegd tijdens mijn sollicitatiegesprek in de zomer van 1996. Nog niet eens klaar als basisarts, zou ik mij al moeten vastleggen voor een AGNIO-schap revalidatiegeneeskunde van zes maanden dat – bij gebleken geschiktheid voor het vak – door zou lopen in een AGIKO-traject van 7,5 jaar. Twijfel aan mijn jawoord op die vraag heb ik eigenlijk nooit gehad. Daarvoor moet ik een heleboel mensen bedanken. Dat ga ik nu doen. Mocht ik je vergeten zijn, bel me dan op; ik zal het dan alsnog goed maken.

Beste Guus, jij was verantwoordelijk voor bovengenoemde vraag, maar zorgde er ook voor dat ik nooit ernstig ging twijfelen aan mijn antwoord. De vrijheid die je me gaf om me te ontwikkelen en je grote inzet voor de revalidatiegeneeskunde zijn hierbij voor mij belangrijke factoren geweest. Beste Lex, 'jouw' methodologie, zowel met betrekking tot de epidemiologie als de manuscript-correctiemethode, is onnavolgbaar. Ik heb er erg veel van geleerd, waarvoor mijn dank. Beste Chris, jouw kennis van het vakgebied MS was altijd behulpzaam bij het schrijven van mijn manuscripten. Kleine opmerkingen over balans in de discussie en de klinische boodschap zorgden vervolgens voor uren denkwerk om dat te bereiken.

Beste Heleen, als copromotor ben jij het hele traject zeer intensief betrokken geweest. In elke fase heb ik een hoop van je opgestoken. Je eindeloze uitdagingen om het onderste uit de kan te halen, je nauwgezette lezing van alles wat ik je heb voorgelegd en de dis-

cussies over de te kiezen aanpak hebben onmiskenbaar bijgedragen aan het mooie eindresultaat van die jaren werk! Voor de komende jaren hebben we nog werk genoeg te doen. Ik kijk er naar uit.

Het is ongelooflijk hoeveel mensen een rol hebben gespeeld bij het slagen van dit onderzoeksproject. Zonder de medewerking van de neurologen in het VUmc (Chris Polman, Bernard Uitdehaag, Brechtje Jelles, Bob van Oosten en Jan Meilof), het Erasmus MC (Prof. Dr. F.G.A. van der Meché, Dr. P.A. van Doorn, Rogier Hintzen), het AMC (Prof. Dr. M. Vermeulen), het Sint-Lucas Andreas ziekenhuis (Dr. J.A.L. Vanneste), en het OLVG ziekenhuis (Drs. H.K. van Walbeek) hadden we nooit 156 patiënten kunnen includeren. Op deze plaats allemaal bedankt voor jullie medewerking.

Zonder deelnemers aan onderzoek komen er geen nieuwe wetenschappelijk inzichten. Alle deelnemers aan dit onderzoek wil ik daarom hartelijk bedanken voor het herhaaldelijk invullen van die lange vragenlijst, het doen van alle testen – ook die vervelende sommetjestest – en het laten maken van de MRI-scans.

En dan heb je de deelnemers die mee willen doen, maar hoe ontfutsel je ze hun geheimen die van belang zijn voor het beantwoorden van de vraagstellingen van het onderzoek? Dit is een delicate aangelegenheid. De gegevens moeten nauwkeurig en compleet verzameld worden, terwijl de belasting van de patiënt ook in de gaten gehouden moet worden. Monique, Maaike, en Tessa, jullie bleken uitstekend in staat om deze balans te vinden. BEDANKT!

Iedereen die betrokken geweest is bij het plannen, maken, beoordelen en scoren van de MRI-scans (early 11) wil ik op deze plaats bedanken.

Het analyseren van de gegevens en het schrijven van de artikelen is een multidisciplinaire aangelegenheid en is niet mogelijk zonder inhoudsdeskundige en kritische coauteurs. Dankzij hun werk wordt de inhoud beter en worden de formuleringen scherper. Bernard, Riekie, Jos, Arian, en Martijn, ik werk met veel plezier met jullie samen.

De leescommissie bestaande uit Prof. Dr. J. Dekker, Dr. R.Q. Hintzen, Prof. Dr. M.A.G. Sprangers, Prof. Dr. E. Lindeman, en Prof. Dr. G.A.M. van den Bos wil ik graag bedanken voor hun nauwgezette beoordeling van het proefschrift.

Bij het opstarten van het onderzoek heb ik veel gehad aan mijn kamergenoten, Eline (Roelofsen) en Hanneke, en mijn voorganger, Leo. Met veel plezier denk ik terug aan de discussies van uiteenlopende aard die we gevoerd hebben. Cecile, ondanks de afstand heb ik altijd prettig met je samengewerkt.

Naast het uitvoeren van het onderzoek moest ik ook nog tot revalidatiearts opgeleid worden. De eerste twee jaar van de opleiding bracht ik in Heliomare door. Rob, jij hebt mij wegwijs gemaakt in de klinische revalidatie. Hoewel ik je soms fronsend heb aangekeken, heb ik wel veel van je geleerd. Pauline en Coen, jullie waren minder intensief bij mijn opleiding betrokken, maar hebben desondanks met enthousiasme de kennis over de wondere wereld van de cognitieve revalidatie overgebracht. Hans, als opleider heb je me veel ruimte gegeven om onderzoeksverplichtingen en opleidingsverplichtingen met elkaar te combineren. Daarnaast heb ik veel waardering voor je inzet voor de opleiding.

De laatste twee jaar van de opleiding vonden plaats in het VUmc. Jules, Frans, Joke (ten Cate), Arianne, Carolien, Mirjam en Joke (van der Neut), ieder van jullie wil ik hartelijk danken voor de goede opleidingssfeer en het overbrengen van de kennis die jullie hebben van jullie aandachtsgebied. Beste Jules, jouw rol begint eigenlijk al eerder. Door een wetenschappelijke stage bij jou werd mijn voorkeur voor de revalidatiegeneeskunde bevestigd. Gelukkig heb jij me tijdens het keuze-co-schap gewezen op de AGIKO-vacature.

Mijn AGIKO-opleiding kreeg natuurlijk mede vorm door de collega AGIO's, collega AGIKO's en andere onderzoekers. Aan het afzien met het clubje hardlopende MS-onderzoekers denk ik nog vaak met plezier terug. Allemaal bedankt voor de leuke tijd.

De leden van de FuPro-groep, Eline (Lindeman), Henk, Joost, Trudi, Anita, Anne, Annet, Gerard, Gert, Marcel, Marjolein, Agnes, Bianca, Imelda, Ingrid en Vera, wil ik graag bedanken voor de betrokkenheid bij het project, de discussies tijdens de vergaderingen en de onderlinge samenwerking.

Alle medewerkers van de afdeling revalidatiegeneeskunde, de leden van het MS-team, het ITB-team en het plexus brachialis-team, zorgen voor een inspirerende werkomgeving waarin ik met veel plezier werk. Monique en Vicky, bedankt voor jullie inzet waar nodig.

Beste Claudia en Erwin, bedankt dat jullie mij vandaag willen bijstaan. Beste Claudia, al vanaf het eerste jaar van de opleiding voeren wij samen practica uit, wie had toen gedacht dat dit zou uitmonden in deze dag. Beste Erwin, wie had dit 14 jaar geleden gedacht... Reuze bedankt dat je het aandurfde om paranimf te worden.

Naast werk zijn er tal van andere leuken dingen in het leven, waar vrienden een belangrijke rol in spelen. Allemaal bedankt. Sporters onder jullie, er zijn alweer plannen om weer met hardlopen te beginnen...

Pa en Ma, jullie hebben er met zijn tweeën toch maar mooi voor gezorgd dat dit allemaal mogelijk is geworden. Heel veel dank daarvoor. Pa, jammer dat je dit niet meer mee kunt maken. Sjaak en Lies, bedankt voor jullie interesse in mijn werk en jullie hulp waar nodig.

Lieve Wendy, Jurre en Thije, veel woorden heb ik niet nodig, ik hou van jullie.

Publications

2007 Dallmeijer AJ, de Groot V, Roorda LD, Schepers VP, Lindeman E, van den Berg LH, Beelen A, Dekker J; FuPro Study Group. Cross-diagnostic validity of the SF-36 physical functioning scale in patients with stroke, multiple sclerosis and amyotrophic lateral sclerosis: a study using Rasch analysis. J Rehabil Med 39:163-9.

Jasperse B, Minneboo A, de Groot V, Kalkers NF, van Helden PE, Uitdehaag BMJ, Barkhof F, Polman CH. Determinants of cerebral atrophy rate at the time of diagnosis of multiple sclerosis. Arch Neurol 64:190-4.

2006 Dallmeijer AJ, Dekker J, Knol DL, Kalmijn S, Schepers VP, de Groot V, Lindeman E, Beelen A, Lankhorst GJ; on behalf of the FuPro study group. Dimensional structure of the SF-36 in neurological patients. J Clin Epidemiol 59:541-3.

de Groot V, Beckerman H, Uitdehaag BMJ, de Vet HC, Lankhorst GJ, Polman CH, Bouter LM. The usefulness of evaluative outcome measures in patients with multiple sclerosis. Brain 129:2648-59.

Huijbregts SCJ, Kalkers NF, de Sonneville LMJ, de Groot V, Polman CH. Cognitive impairment and decline in different MS subtypes. J Clin Sci 245:187-94.

2005 Dallmeijer AJ, Dekker J, Roorda LD, Knol DL, van Baalen B, de Groot V, Schepers VP, Lankhorst GJ. Differential item functioning of the Functional Independence Measure in higher performing neurological patients. J Rehabil Med 37:346-52.

de Groot V, Beckerman H, Lankhorst GJ, Polman CH, Bouter LM. The initial course of daily functioning in multiple sclerosis: a three-year follow-up study. Mult Scler 11:713-8.

2004 Huijbregts SCJ, Kalkers NF, de Sonneville LMJ, de Groot V, Reuling IEW, Polman CH. Differences in cognitive impairment of relapsing remitting, secondary, and primary progressive MS. Neurology 63; 335-9.

Bot JC, Barkhof F, Polman CH, Lycklama a Nijeholt GJ, de Groot V, Bergers E, Ader HJ, Castelijns JA. Spinal cord abnormalities in recently diagnosed MS patients: added value of spinal MRI examination. Neurology 62:226-33.

de Groot V, Beckerman H, Lankhorst GJ, Bouter LM. How to measure comorbidity: a critical review of available methods. J Clin Epidemiol 57:323. (Authors reply)

2003 de Groot V, Beckerman H, Lankhorst GJ, Bouter LM. How to measure comorbidity: a critical review of available methods. J Clin Epidemiol 56:221-9.

2002 Uitdehaag BMJ, Ader HF, Roosma TJA, de Groot V, Kalkers NF, Polman CH. Multiple Sclerosis: impact of reference population and interpretation of changes. Mult Scler 8:366-71.

2001 van der Lee JH, de Groot V, Beckerman H, Wagenaar RC, Lankhorst GJ, Bouter LM. The intra- and interrater reliability of the action research arm test: a practical test of upper extremity function in patients with stroke. Arch Phys Med Rehabil 82:14-9.

Kalkers NF, Bergers E, de Groot V, Lazeron RH, van Walderveen MA, Uitdehaag BMJ, Polman CH, Barkhof F. Concurrent validity of the MS Functional Composite using MRI as a biological disease marker. Neurology 56:215-9.

2000 Kalkers NF, de Groot V, Lazeron RH, Killestein J, Ader HJ, Barkhof F, Lankhorst GJ, Polman CH. MS functional composite: relation to disease phenotype and disability strata. Neurology 54:1233-9.

About the author

Vincent de Groot was born on the March 9, 1971 in Alkmaar. In 1989 he graduaded from secondary school at the Oscar Romero college in Hoorn. Between 1989 and 1996 he studied medicine at the VU University in Amsterdam. After he received his medical degree in November 1996, he started as medical assistant at the department of rehabilitation medicine at the VU University Medical Center. Between June 1997 and December 2004 he continued to work as a Medical Trainee for Clinical Research (AGIKO) at the same department and at Heliomare rehabilitation center in Wijk aan Zee. In this period he carried out the research for this thesis, followed the post-graduate course epidemiology at the EMGO Institute, and was trained as a physiatrist. Vincent de Groot is currently working as a physiatrist and researcher at the department of rehabilitation medicine at the VU University Medical Center. His areas of interest include rehabilitation for neurological disorders and obstetric brachial plexus palsy.

Vincent de Groot werd geboren op 9 maart 1971 te Alkmaar. In 1989 behaalde hij het VWO-examen aan de scholengemeenschap Oscar Romero te Hoorn. Van 1989 tot 1996 studeerde hij geneeskunde aan de VU in Amsterdam. Direct na het behalen van zijn arts-examen in november 1996, startte hij als Assistent Geneeskundige Niet In Opleiding (AGNIO) revalidatiegeneeskunde in het VUmc. Van juni 1997 tot december 2004 werkte hij als Assistent Geneeskundige In opleiding tot Klinisch Onderzoeker (AGIKO) revalidatiegeneeskunde

op dezelfde afdeling en in het revalidatiecentrum Heliomare te Wijk aan Zee. Tijdens deze periode werd het onderzoek voor dit proefschrift uitgevoerd, volgde hij de postdoctorale opleiding epidemiologie aan het Instituut voor Extramuraal Geneeskundig Onderzoek (EMGO), en werd hij opgeleid tot revalidatiearts. Van 2004 tot heden werkt hij als revalidatiearts en onderzoeker in het VUmc met als aandachtsgebieden revalidatie bij neurologische aandoeningen en het obstetrisch plexus brachialis-letsel.