Hellriegel, Barbara; Daumer, Martin; Neiss, Albrecht

Working Paper

Analysing the course of multiple sclerosis with segmented regression models

Discussion Paper, No. 355

Provided in Cooperation with:
Collaborative Research Center (SFB) 386: Statistical Analysis of discrete structures - Applications in Biometrics and Econometrics, University of Munich (LMU)

Suggested Citation: Hellriegel, Barbara; Daumer, Martin; Neiss, Albrecht (2003) : Analysing the course of multiple sclerosis with segmented regression models, Discussion Paper, No. 355, Ludwig-Maximilians-Universität München, Sonderforschungsbereich 386 - Statistische Analyse diskreter Strukturen, München,
https://doi.org/10.5282/ubm/epub.1730

This Version is available at:
https://hdl.handle.net/10419/31091

Discussion paper 355 - SFB 386 - LMU Munich

Analysing the course of multiple sclerosis with segmented regression models.

Barbara Hellriegel[1,2], Martin Daumer[1,3], and Albrecht Neiß[1,2]

[1] Sylvia Lawry Centre for MS Research

[2] Institute for Medical Statistics and Epidemiology

[3] Trium Analysis Online

August 2003

Abstract

Multiple sclerosis (MS) is a demyelinating disease of the central nervous system whose cause is still unknown. The disease course shows great inter- and intra-individual variability and this results in insecurity of diagnosis and prognosis. A well-founded knowledge of the natural history of MS, however, is an important prerequisite for developing adequate strategies for therapy and research. In order to increase our understanding we developed a segmented regression model which extracts three main characteristics of the time course of this complex disease from natural history data. For each individual patient this model determines baseline disability (as measured by the Expanded Disability Status Scale = EDSS), the time point where the disease starts to progress and the slope of this progression. The model is applied to data of patient registries from all over the world that are pooled in the database of the Sylvia Lawry Centre for Multiple Sclerosis Research (SLCMSR). The analyses used a random subsample of the entire database and were restricted to patients seen from onset of MS with time series of at least three years. Thereby we were able to avoid some of the problems related to missing data. Our results revealed a weak negative correlation between time to progression (change point) and slope of progression for this group of patients, i.e. those patients who do progressed later and remained stable for a longer time developed disability more slowly than those who progressed earlier. For the two parameters and their interaction we did not find an influence of basic covariates like gender, disease course and mono- or poly-symptomatic disease onset. According to the SLCMSR Policy these results will be subjected to a validation using an independent "validation dataset". This remains to be done.

1 Introduction

Multiple sclerosis (MS) is a demyelinating disease of the central nervous system. It is influenced by genetic and environmental factors but its cause is still unknown. Considerable research has revealed a great and to a large extent unexplained heterogeneity in disease course which results in insecurity of diagnosis and prognosis.

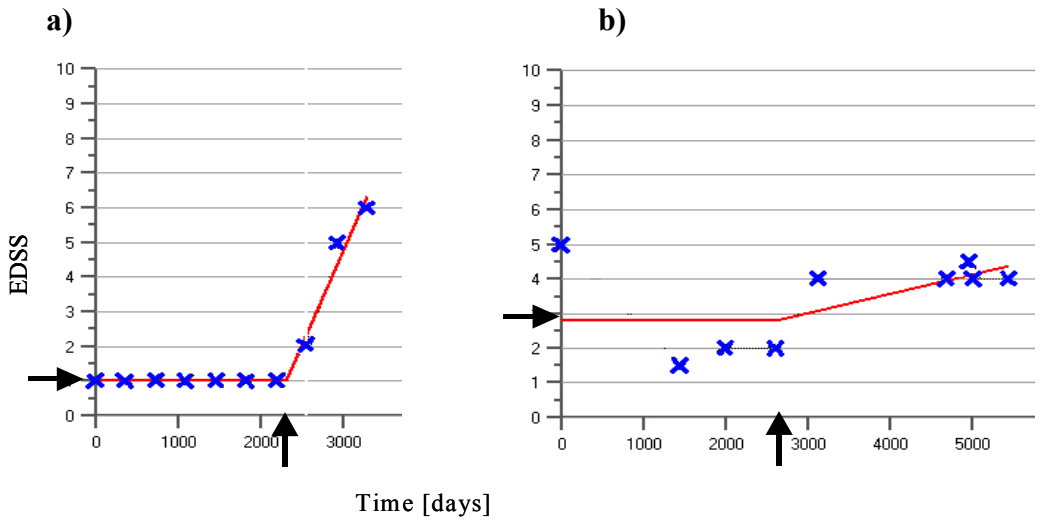
A good understanding of the natural course of MS is, however, an important basis for developing adequate strategies for therapy and research. The long-term evolution of MS in an individual patient usually shows one or two phases. After an initial, more or less stable phase (which can also be absent) disability increases steadily through the accumulation of partially reversible impairments of the central nervous system. The change point between stable and progressive phase is usually not easily determined.

Analyses of the natural course of MS are usually performed using survival methods for the time to reach sustained progression or critical disability levels (e.g. Weinshenker et al., 1991; Amato et al., 1999; Confavreux et al., 2003). Several of the older articles find themselves in disagreement over the influence of important covariates such as gender, age at disease onset, initial symptoms, type of disease course etc. This is partly due to the fact that different endpoints/outcome measures were considered but also to the use of statistical methods which did not adequately account for interdependencies between potential factors of influence. More recent work using comparable endpoints and more complex (e.g. multivariate) statistical methods draw similar conclusions with respect to the influence of some basic predictors (gender, initial symptoms, type of disease course), despite using different data sets. The importance of other covariates, such as the number of relapses in a given time interval, is still equivocal (e.g. Weinshenker et al., 1991; Confavreux et al. 2000). A Markov model makes clear that the effects of some covariates are limited to transitions between specific disease states (Wolfson and Confavreux, 1987). As all of these results only allow for very rough predictions for categories of patients with remaining high variation within them the challenge is now to advance existing and develop new methods enabling more specific forecasts at early disease stages. The approach taken here is a step in this direction.

Data for individual MS patients usually consist of time series of neurological measures (see Figure 1) implying that survival models neglect valuable information. Nevertheless, methods for longitudinal data are rarely applied. Fog and Linneman (1970) and Patzold and Pocklington (1982) use regression analyses to choose the best fitting among several concurrent regression curves (linear, parabolic, hyperbolic and polynomial). Both studies arrive at similar conclusions although their time series are based on different neurological scales. On the other hand they result in categories of regression curves which are difficult to compare. This also complicates further analyses. In order to escape this dilemma we developed a segmented regression model which extracts three main characteristics of the time course of disability from natural history data. For each patient this model determines baseline disability, the time point where progression starts and the slope of this progression. Disability is measured using

the Expanded Disability Status Scale (=EDSS), an ordinal scale ranging from zero to ten with steps of 0.5. The data stem from a large pooled database of registry data which have been donated to the Sylvia Lawry Centre for Multiple Sclerosis Research (SLCMSR) by institutions from all over the world (see acknowledgement; Noseworthy et al. 2003). Methodologically our approach ties in with models for detecting change points (e.g. Basseville and Nikiforov 1993; Daumer and Neiß 2001) and as regards interpretation it is based on widespread assumptions on the underlying physiological process (see above). In order to reduce potential difficulties arising from missing values and measurement error we restricted the current analysis to patients seen from onset of MS with time series of at least three years and three or more visits.

Figure 1 Examples of individual time series of disability as measured by the Extended Disability Status Scale (EDSS) and model output (red line). The horizontal arrows indicate the baseline EDSS (e_0) and the vertical arrows the change point (τ) determined by the model.



2 Database and patient population

The data have been donated to the SLCMSR from institutions all over the world (see acknowledgements). For confidentiality reasons it is not possible to identify the individual sources of the data pooled in the SLCMSR database. Moreover, to warrant a high quality of statistical analyses performed at the SLCMSR, the database is randomly split into two subsets: the open part used for model building and hypotheses generation and the closed part is administered by trustees. It is the declared policy of the SLCMSR to validate analyses performed on the open part of the database using its closed counterpart. The results presented here have not yet undergone this validation procedure and have, therefore, to be considered as preliminary.

The analyses in this study utilized a random subset of the natural history part of the SLCMSR database which in release 003 contains 1628 patients with more than one visit. Information was available on the following covariates: gender; age at onset of disease [years]; mono- or poly-symptomatic disease onset; duration of MS [months], age and EDSS at entry into the study; and disease course (RR: relapsing remitting, SP: secondary progressive, RP: relapsing progressive, PP: primary progressive).

The focus of this study were patients seen at onset (N=167), i.e. whose first presentation coincided with the onset of MS, and who additionally qualified as valid for the analyses. Initially we included all those patients with three or more visits and a time series of at least three years after onset (N=93). In order to account for the fact that the status five years after onset is an important predictor (e.g. Kurtzke et al. 1977), the sample was further reduced after fitting the segmented regression model described below. Patients who were censored before the sixth year while still being stable were removed from the sample because their status (stable or progressing) at year five was unknown. In what follows the resulting group of 62 patients will be termed valid. Two thirds of these patients were female (see Table 1) and 40.3% had a mono-symptomatic onset. The disease course was relapsing remitting in 29.0% of the cases, 51.6% were secondary progressive, 8.1% relapsing progressive, one woman (1.6%) had a primary progressive course and for 9.7% the information was missing.

3 Modelling approach and statistical analyses

In order to capture the basic characteristics of the disease course a segmented regression model with change-point τ was chosen:

$$E_i(Y \mid X = x) = e_0 + a\,(x-\tau)_+ \quad mit \quad t_+ = \begin{cases} t \ \textit{für}\ t \geq 0 \\ 0 \ \textit{für}\ t < 0 \end{cases} \quad und \quad i = 1,..,N$$

Our approach restricted the slope of the first regression line to be equal to zero (α_0=0) and that of the second to be positive (α>0). Thus, for each individual time series (i=1, .., N) three clinically important characteristics were determined: (i) the baseline EDSS e_0 in the first segment, (ii) the time point τ where progression starts (change point) and (iii) the slope α of this progression in the second segment (Figure 1). The ML estimation of these model parameters is equivalent to a non-linear L^2 approximation problem and can be solved analytically (cf. Küchenhoff, 1997). With n data points of a times series and k possible change points there exists a maximum of ($O(n^k)$) local

minima which can be determined explicitly. The algorithm determining the model parameters was programmed using ActivePerl (1996-2002).

These model parameters and their relation were then subjected to statistical analyses. The influence of covariates on α and τ, respectively, was investigated using multiple regression analyses and the simultaneous influence on both parameters was studied with multivariate methods. Logistic regression was employed to identify differences between subgroups of patients. Statistical analyses were conducted using SPLUS 6.0 (1988-2000) and SPSS 11.0 (2001) both for Windows.

4 Results

As can be seen from Table 1, the 167 patients seen at onset developed MS at an earlier age and tended to have a more severe course than those seen 4.3±6.7 years (mean±SD, N=1461) after the onset of MS. The 62 valid patients who were used for further analyses form a representative subset (Table 1).

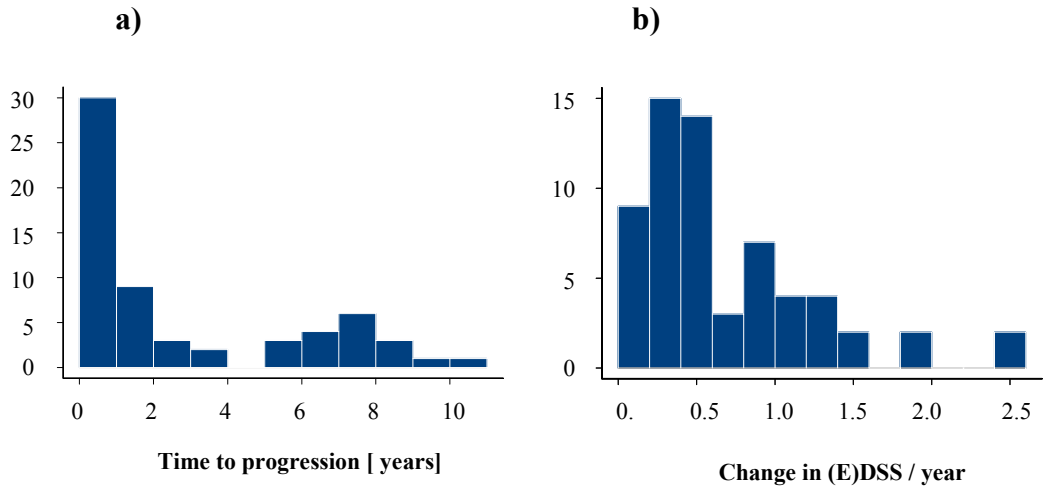
Table 1. Characteristics of study population. Patients seen at onset of MS are compared to those seen only later (*t*-test). Sixty-two patients seen at onset proofed valid for further analyses (for details see "Database and patient population").

	Seen at onset		**Seen later**	**P-value**
Number – total	167	62	1461	--
Women	71.3% (119)	67.7% (42)	64.6% (944)	--
Age at onset of MS				
- mean ± SD [years]	28.3 ± 9.7	28.6 ± 9.4	30.9 ± 10.0	0.0009
- median, range [y]	26, 12-54	27, 12-54	30, 6-72	
Age at 1st presentation				
- mean ± SD [years]	28.3 ± 9.7	28.6 ± 9.4	35.8 ± 11.4	< 0.0001
- median, range [y]	26, 12-54	27, 12-54	35, 8-82	
(E)DSS at 1st present.				
- mean ± SD	2.2 ± 1.5	2.3 ± 1.5	2.5 ± 1.7	0.029
- median, range	2.0, 1-8	2.0, 1-8	2.0, 0-8.5	

The fact that the distribution of baseline EDSS values (e_0) calculated by the model had very similar moments to that of the EDSS at first presentation (mean±SD: 2.2 ± 1.5, median: 2.0) was reassuring. Assuming that at least some first EDSS values resulted from relapses it was not unexpected to find a smaller range for the baseline EDSS (1-6,

cf. Table 1) which was determined using more than just the first value. The distribution of the time to progression (change point) seems to have to modes (Figure 2a). The forty-four patients (71%) who progressed early did so before the fifth year after onset of MS (range:0-4), most of them even during the first year (71.4%, median: 0, mean±SD: 0.8±1.1 years). The late group progressed after a mean of 7.6±1.4 (±SD, median: 7.1, range: 5.6-11). The yearly change in EDSS (slope of disease progression) showed a distribution which was highly skewed to the right (median: 0.48, mean ±SD: 0.67±0.58, range: 0-2.5; Figure 2b), with 56.5% of the patients changing by less than 0.5 per year on the EDSS scale.

Figure 2 Histograms of a) time to start of disease progression (change point) and b) change in EDSS per year as determined by the model.

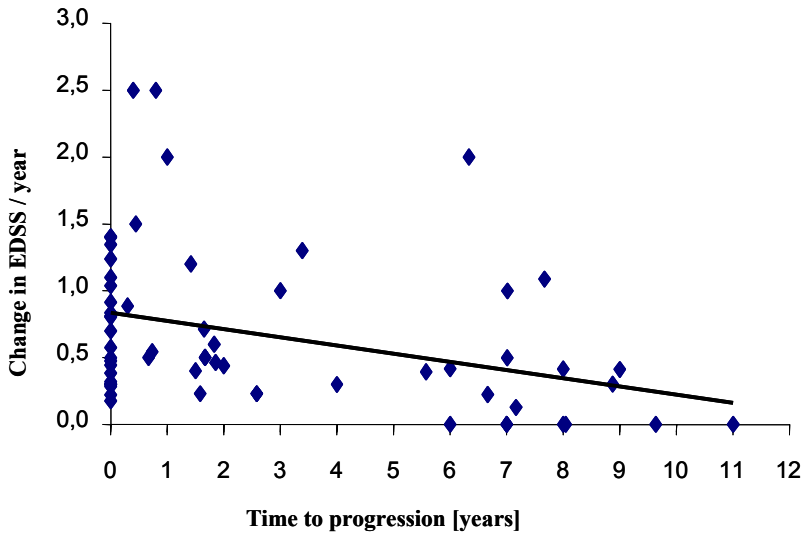


The correlation between time to progression (τ) and yearly EDSS change (α) was weak and negative (correlation coefficient of Spearman ρ=-0.36, p=0.0046, Figure 3). For the subset of 32 patients with a two-phase course ($\tau \neq 0$ and $\alpha \neq 0$) the negative correlation was slightly higher (ρ=-0.44, p=0.014, Y=2.7-0.3x, R^2=0.17). While the negative correlation was also slightly increased for women (ρ=-0.46, p=0.003, N=42), patients with relapsing remitting (ρ=-0.44, p=0.005, N=18) and secondary progressive course (ρ =-0.42, p <0.02, N=32) or with mono-symptomatic onset (ρ =-0.43, p<0.01, N=37), it disappeared for men (ρ =-0.06, p=0.78, N=20) and patients with poly-symptomatic onset (ρ =-0.28, p=0.17, N=25).

However, when the influence of covariates on α and τ respectively was investigated using multiple regression analyses and when their simultaneous influence on both parameters was studied with multivariate methods we did not find an influence of any of these covariates. When searching for differences between those patients progressing

from onset (τ=0) and later (τ>0) using logistic regression analysis we as well did not discover a dependence on the available covariates.

Figure 3 Relation between time to start of disease progression (change point) and change EDSS per year as determined by the model (N=62; Y=0.8-0.06x, R^2=0.12).



5 Discussion

The time point when MS patients with a non-progressive onset proceed from an initial, more or less stable to a progressive phase is usually not easily determined. This hinders classifying patients into groups of those remaining stable or progressing either immediately or beyond specific time points after onset and using this classification, for instance, for making predictions. We therefore developed a model which determines this transition point together with the patient's baseline EDSS level before and the slope of disease progression afterwards. This model was then fitted to the individual time series of EDSS values of patients seen from disease onset. As the EDSS value five years from onset is known to be a predictor of the further development of disability (e.g. according to Kurtzke's five year rule it on average amounts to three quarters of that at year 15; Kurtzke et al. 1977) we determined the patients' status every five years. Five years after onset eighteen (29%) of the 62 patients seen at onset were still stable, 33.9% had progressed after an initial stable phase and 37.1% had progressed immediately. Another five years later, only one of the eighteen stable patients had remained stable and 13 had progressed (four stable patients were censored without progressing). That is, after ten years 1.6% of the original 62 patients were still stable

and 91.9% had progressed. This, however, says nothing about the rate of progression of these patients.

In order to see whether the time point when progression starts influences its rate we investigated the relation between time to progression and change in EDSS per year. The results showed a week but significant negative correlation between these two parameters for patients seen at disease onset. This implies that those patients who did progress later and remained stable for a longer time developed disability more slowly than those who progressed earlier. With respect to our goal of facilitating predictions, none of our uni- and multivariate analyses found any support for the influence of basic covariates like gender, mono- or poly-symptomatic onset and disease course on time point and rate of progression or on their interaction. This was not totally unexpected because Fog and Linneman (1970) and Patzold and Pocklington (1982) who also apply regression analyses to time series of neurological measures report similar findings.

The results presented here are preliminary in two ways. Firstly, they have not yet undergone the validation procedure implemented by the SLCMSR to ensure quality standards (see "Database and patient population"). Secondly, the parameter values determined by fitting the segmented regression model are, of course, sensitive to the noisiness of the data. EDSS time series, especially those from registry databases, are known to be influenced by acute attacks which do not reflect the long-term course of disability, by inter- and intra-rater variability of measurements and by inconsistencies of the EDSS scale itself. Possible effects of this are illustrated in Figure 1b where the patient probably presented to the hospital because of a severe relapse (EDSS 5). Taking into account the further course, the model determined a baseline EDSS value of about three (see arrow, Figure 1b). A clinician, on inspection, may assign a baseline value of two. These considerations suggest that it may be worthwhile to subject the time series of EDSS values to a pre-editing procedure which implements neurological prior knowledge to reduce the unavoidable noise in the data. This in turn would allow to make more effective use of the modeling approach described here.

Acknowledgements

We thank C. Confavreux, G. Ebers, S. Kessner, C. Polman, S. Vukusic for discussions and comments, C. Lederer for programming in Perl, and L. Kappos for medical advice. The SLCMSR thanks all its data donors which are listed at www.slcmsr.org/en/partner.htm. BH is funded by the SFB 386 of the German Science Foundation and the SLCMSR by the Multiple Sclerosis International Federation (MSIF).

References

Amato, M.P., Ponziani, G., Bartolozzi, M.L. and Siracusa, G. J. (1999). A prospective study on the natural history of multiple sclerosis: clues to the conduct and interpretation of clinical trials. *Journal of the Neurological Sciences* 168, 96-106.

Basseville, M. and Nikiforov, I.V. (1993): *Detection of Abrupt Changes*, Prentice Hall, Englewood Cliffs, New Jersey.

Confavreux, C., Vukusic, S., and Adeleine, P. (2003). Early clinical predictors and progression of irreversible disability in multiple sclerosis: an amnesic process. *Brain* 126, 770-782.

Confavreux, C., Vukusic, S., Moreau, T. and Adeleine, P. (2000). Relapses and progression of disability in multiple sclerosis. *N. Engl. J. Med*. 343, 1430-1438.

Daumer, M. and Neiss, A. (2001) A new adaptive algorithm to detect shifts, drifts and outliers in biomedical time series. In: J. Kunert, G. Trenkler (eds.) Mathematical Statistics with Applications in Biometry, Josef Eul, Lohmar, pp. 201-204.

Di Serio, C. and Lamina, C. (2003). Bayesian P-Spline to investigate the impact of covariates on Multiple Sclerosis clinical course. *Discussion Paper* 353, SFB 386, Ludwig-Maximilians-Universität München.

Fog, T. and Linnemann, F. (1970). The course of multiple sclerosisin 73 computer designed curves. *Acta Neurol. Scand*. 46 [Suppl. 47], 1-175.

Küchenhoff, H. (1997). An exact algorithm for estimating breakpoints in segmented generalized linear models. *Computational Statistics* 12, 235-247

Kurtzke, J.F., Beebe, G.W., Nagler, B., Kurland, L.T. and Auth, T.L.., 1977. Studies on the natural history of multiple sclerosis 8: Early prognostic features of the later course of the illness. J. Chron. Dis. 30, 819-830.

Noseworthy, J., Kappos, L. and Daumer, M. (2003). Competing interests in multiple sclerosis research. Lancet 361, 350-351.

Patzold, U. and Pocklington, P.R. (1982). Course of multiple sclerosis. First results of a prospective study out of 102 MS patients from 1976-1980. *Acta Neurol. Scand*. 65, 248-266.

Weinshenker, B.G., Bass, B., Rice, G.P.A., Noseworthy, J., Carriere, W., Baskerville J. and Ebers, G.C. (1991). The natural history of multiple sclerosis: a geographically based study. 3. Multivariate analysis of predictive factors and models of outcome. *Brain* 114, 1045-1056.

Wolfson, C. and Confavreux, C. (1987). Improvements to a simple Markov model of the natural history of multiple sclerosis. *Neuroepidemiology* 6, 101-115.