Datenabhängiges, modellbasiertes Gruppieren von binären longitudinalen Verläufen am Beispiel der Neurodermitis

Oliver Kuß*; Cora Gromann*; Thomas L. Diepgen**;

*Institut für Medizinische Epidemiologie,
Biometrie und Informatik,
Martin-Luther-Universität Halle-Wittenberg, Halle(Saale)

** Abteilung Klinische Sozialmedizin, Medizinische
Universitätsklinik Heidelberg

GMDS 2003, Münster, 15.9.2003

Programm

- Neurodermitis
- Daten
- Modell
- Ergebnis
- Diskussion

Neurodermitis

Neurodermitis ist eine Hautkrankheit, die bevorzugt im Kindesalter auftritt, schubhaft verläuft, eine starke genetische Komponente hat und sich klinisch durch Hautekzeme mit quälendem Juckreiz zeigt (Jung, 2003).

Die Prävalenz der Neurodermitis steigt an, was mit einer Veränderung von Umwelt- und Lebensstilfaktoren zusammenhängt.

Die Ätiologie der Neurodermitis ist noch nicht ganz geklärt, Prognosen über den Verlauf sind schwierig.

Eine Klassifizierung von Krankheitsverläufen liegt noch nicht vor, wäre aber (1) aus Präventionsgründen und (2) zur Generierung neuer Hypothesen über Ätiologie und Pathogenese hilfreich.

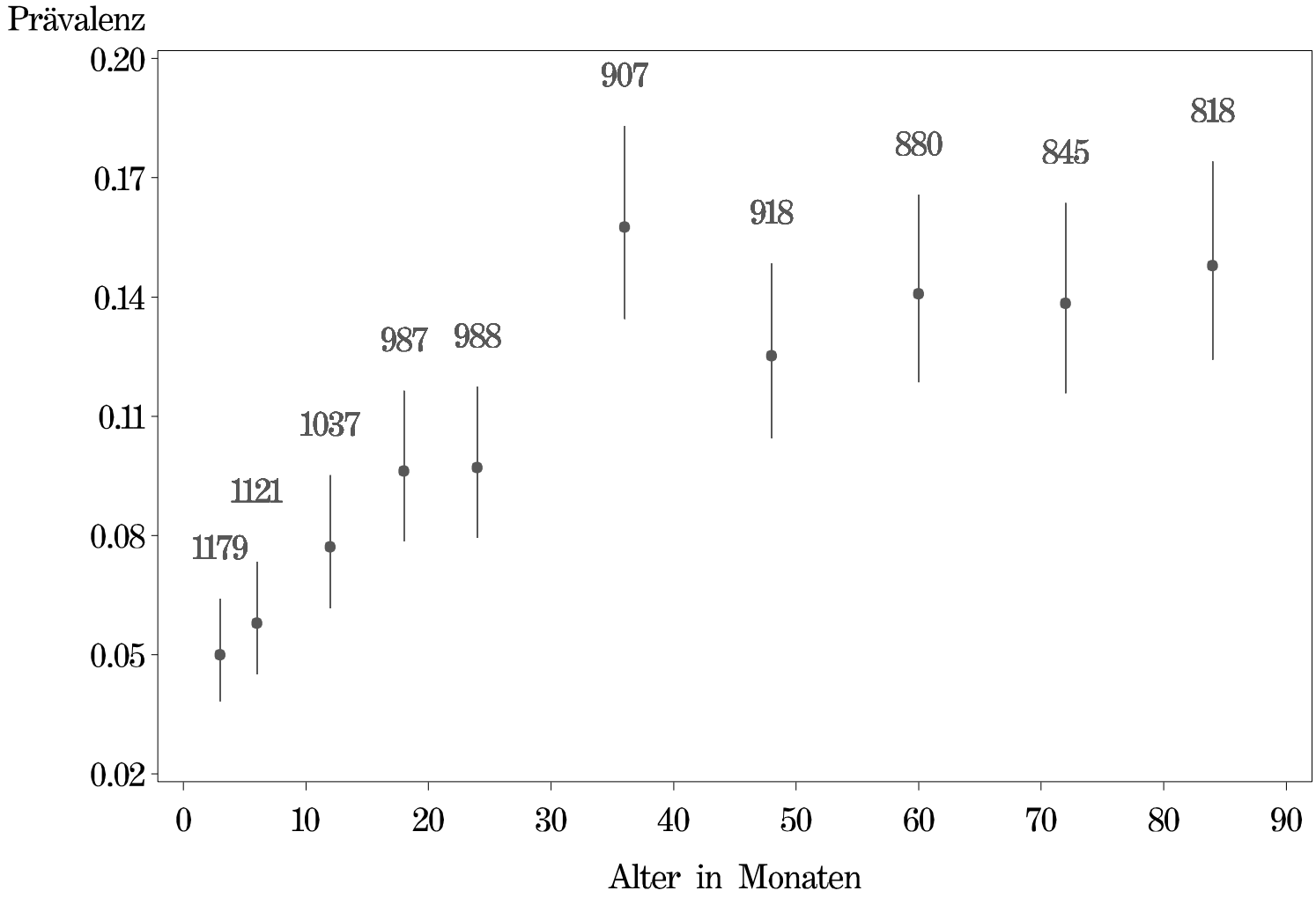
Daten

MAS(= Multizentrische Atopie-Studie)

Prospektive Kohorten-Studie von 1990-1997 in 6 Geburtsstationen in fünf deutschen Städten

1314 Kinder (von ursprünglich 1986 eingeladenen, 499 aus einer „Risikogruppe“)

Follow-Up bis zum siebten Lebensjahr, Erhebung von Neurodermitis-Symptomen und Risikofaktoren an zehn Zeitpunkten (3, 6, 12, 18, 24, 36, 48, 60, 72, 84 Monate nach der Geburt)


Prävalenz
0.20
0.17
0.14
0.11
0.08
0.05
0.02
1179
1121
1037
987
988
907
918
880
845
818
0
10
20
30
40
50
60
70
80
90
Alter in Monaten

Das Modell I

Grundlegende Annahme: In der Population liegen unterschiedliche, nicht beobachtbare (latente) Gruppen von Krankheitsverläufen vor. ($\rightarrow$ Latent Class Mixture Model, Nagin, 1999)

Notation: Gegeben sind Beobachtungen von $I(i = 1, \ldots, I)$ Kindern an $T(t = 1, \ldots, T)$ verschiedenen Zeitpunkten und es gibt $J(j = 1, \ldots, J)$ verschiedene latente Gruppen.

Das Modell II

Die Wahrscheinlichkeit $p(Y_i)$, den Krankheitsverlauf Y_i zu beobachten, ist

$$p(Y_i) = \sum_{j=1}^{J} \pi_j p(Y_{it}|j) \tag{1}$$

wobei Y_{it} das Auftreten der Krankheit an Zeitpunkt t, $p(Y_{it}|j)$ die Wahrscheinlichkeit für dieses Auftreten gegeben die Zugehörigkeit zur latenten Gruppe j und π_j die Wahrscheinlichkeit zur Gruppe j zu gehören, bezeichnet ($\sum_{j=1}^{J} \pi_j = 1$).

Das Modell III

Zur Modellierung der Verläufe in Abhängigkeit vom Alter wird ein logistisches Modell benutzt:

$$p(Y_{it} = 1|j) = \frac{\exp(\alpha_0^j + \alpha_1^j Age_{it} + \alpha_2^j Age_{it}^2 + \alpha_3^j Age_{it}^3)}{1 + \exp(\alpha_0^j + \alpha_1^j Age_{it} + \alpha_2^j Age_{it}^2 + \alpha_3^j Age_{it}^3)} \quad (2)$$

mit Age_{it} als dem Alter von Beobachtung i zum Zeitpunkt t.

Das Modell IV

Die Modellierung von Kovariablen $x_i = (x_{i1}, \ldots, x_{iR})$ auf die Gruppenzugehörigkeiten geschieht mit einem multinomialen logistischen Modell:

$$\pi_j(x_i) = \frac{\exp(\beta_0^j + \beta_1^j x_1 + \cdots + \beta_R^j x_R)}{\sum_j \exp(\beta_0^j + \beta_1^j x_1 + \cdots + \beta_R^j x_R)} \tag{3}$$

Annahme/Einschränkung: Kovariablen beeinflussen nur Gruppenzugehörigkeit, nicht die Form des Krankheitsverlaufs.

$$x \longrightarrow j \longrightarrow Y$$

Kovariablen $\longrightarrow$ Latente Gruppe $\longrightarrow$ Verlauf

Modellwahl/Parameterschätzung

Es muss das Modell mit (1) der optimalen Anzahl von Gruppen, (2) der optimalen polynome Ordnung in jeder Gruppe und (3) mit der optimalen Kovariablenmenge ausgewählt werden.

Da LR-Testen in Mischungsmodellen mit Problemen behaftet ist, verwende BIC (Bayes Information Criterion, Schwarz, 1978)

Parameterschätzung mit SAS PROC NLP

Posterior Wahrscheinlichkeiten

Ein wichtiges Ergebnis sind die geschätzten Posterior Wahrscheinlichkeiten für die Gruppenzugehörigkeit:

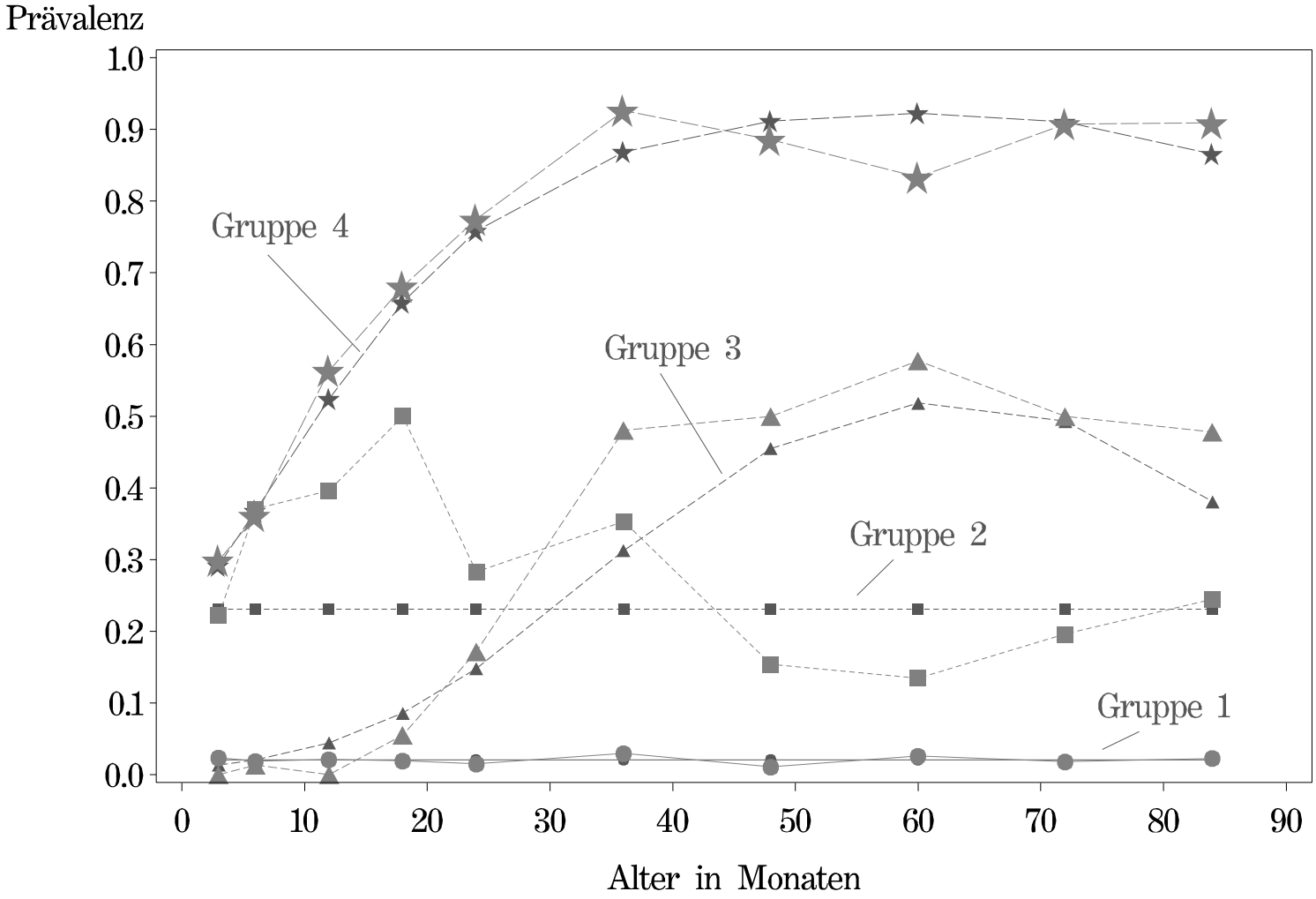
$$\widehat{p}(j|Y_i) = \frac{\widehat{p}(Y_i|j)\widehat{\pi}_j}{\sum_j \widehat{p}(Y_i|j)\widehat{\pi}_j} \tag{4}$$

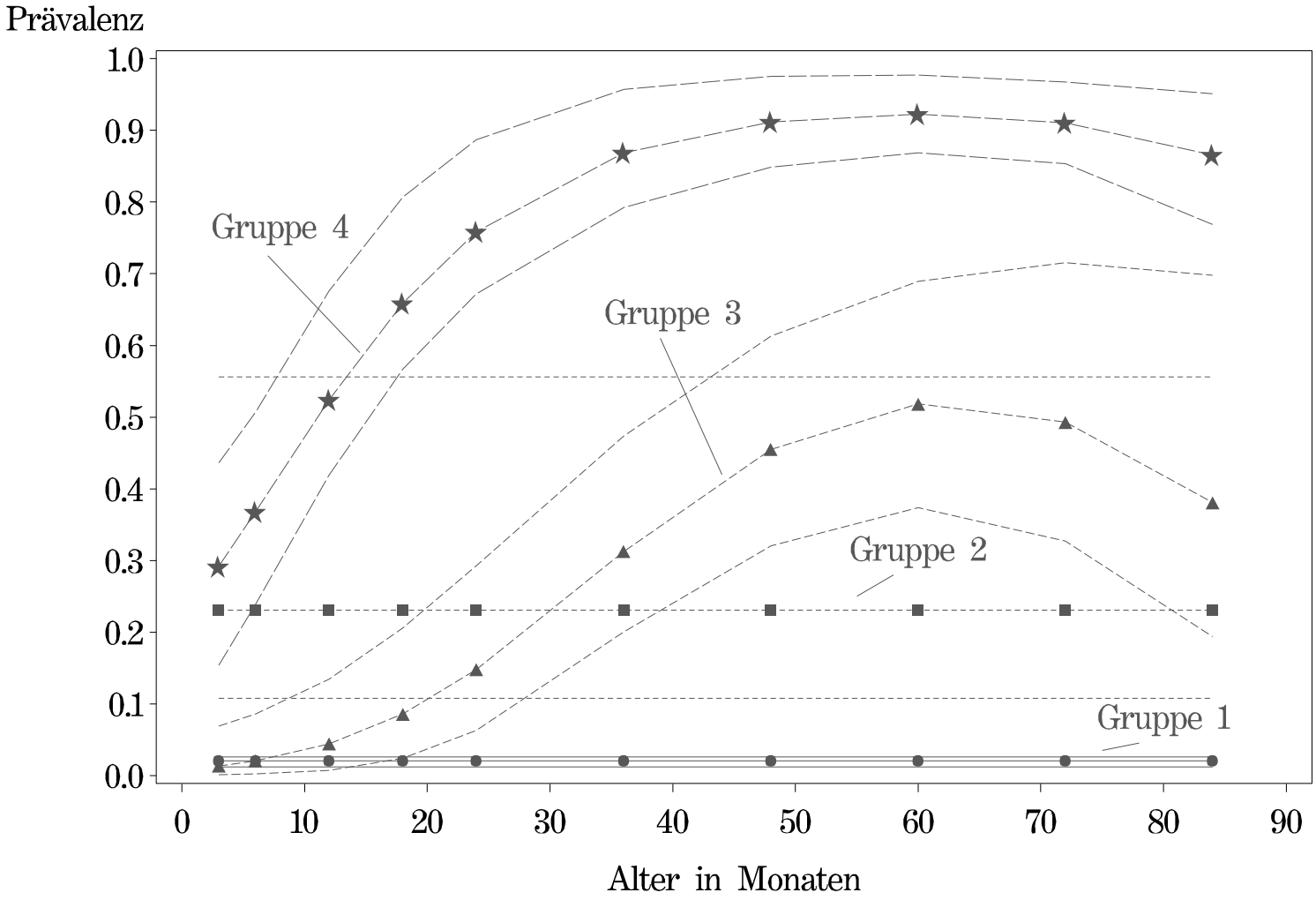
Ermöglicht Zuteilung der Verläufe in Gruppen (maximales $\widehat{p}(j|Y_i)$) und Darstellung von „mittleren“ Verläufen

Ergebnis

Optimales Modell (BIC) hat 4 Gruppen mit 2 Gruppen polynomialer Ordnung und 2 Gruppen konstanter Ordnung, **jedoch keine Kovariablen**.

Gruppe	Parameter	Schätzwert	Parameter	Schätzwert
1	α_0^1	-3.86 (0.15)	π_1	78.8% (2.8%)
2	α_0^2	-1.21 (0.37)	π_2	7.0% (3.0%)
3	α_0^3	-4.75 (0.81)	π_3	8.0% (1.9%)
	α_1^3	1.85 (0.37)		
	α_2^3	-0.18 (0.04)		
4	α_0^4	-1.25 (0.29)	π_4	6.2% (0.9%)
	α_1^4	1.50 (0.27)		
	α_2^4	-0.15 (0.04)		


Prävalenz
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Gruppe 4
Gruppe 3
Gruppe 2
Gruppe 1
0
10
20
30
40
50
60
70
80
90
Alter in Monaten


Prävalenz
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Gruppe 4
Gruppe 3
Gruppe 2
Gruppe 1
0
10
20
30
40
50
60
70
80
90
Alter in Monaten

Diskussion

- Eine sinnvolle, datenabhängige, modellbasierte Gruppierung von Neurodermitis-Krankheitsverläufen konnte identifiziert werden.

- Das Original-Modell von Nagin wurde in zweierlei Hinsicht verallgemeinert: (1) Bootstrap-Konfidenzintervalle für die „mittleren“ Verläufe und (2) Modellierung der Kovariablen mit einem Stereotype-Modell

- Das Modell liefert eine Menge von interpretierbaren und kommunizierbaren Ergebnissen

Literatur

Jung EG, Moll I, [Eds.], (2003): Dermatologie. Stuttgart: Thieme-Verlag.

Williams HC, Strachan DP, (1998): The natural history of childhood eczema: observations from the British 1958 birth cohort study. British Journal of Dermatology, 139, 834-839.

Nagin DS, (1999). Analyzing Developmental Trajectories: A Semi-Parametric, Group-Based Approach. Psychological Methods, 4, 139–177.

Schwarz G, (1978). Estimating dimensions of a model. Annals of Statistics, 6, 461-464.