Data Mining in einer epidemiologischen Studie zu Oberbauchbeschwerden (PRESTO)

Guido Knapp, Joachim Hartung
Universität Dortmund
Fachbereich Statistik
44221 Dortmund
knapp@statistik.uni-dortmund.de
hartung@statistik.uni-dortmund.de

Zusammenfassung

In einer epidemiologischen Studie zu Oberbauchbeschwerden (PRESTO) werden Data Mining Methoden, die im SAS Enterprise Miner zur Verfügung stehen, benutzt, um Prognosemodelle für Studienresponder zu erstellen. Dabei werden Responder sowohl bezüglich der Reduktion der Beschwerden als auch bezüglich der Verbesserung der Lebensqualität betrachtet. Es wird gezeigt, dass mit den Data Mining Methoden sinnvolle prognostische Faktoren für die Responder identifiziert werden können.

Keywords: Funktionelle Dyspepsie, Logistische Regression, Entscheidungsbäume, Künstliche Neuronale Netze, SAS Enterprise Miner^{TM}

1 Problemstellung

Die Darstellung der zugrunde liegenden Problematik orientiert sich im wesentlichen an der Arbeit von Allescher et al. [1].

Die Dyspepsie ist definiert als „chronische oder intermittierende Beschwerden, die dem oberen Verdauungstrakt zugeschrieben werden“. Man unterscheidet zwischen der organischen (zum Beispiel Ulkusleiden) und der idiopathischen bzw. funktionellen Dyspepsie, bei der keine Läsionen feststellbar sind.

In Deutschland beträgt die Prävalenz dieses Symptomenkomplexes immerhin rund 25%.

In der hausärztlichen Praxis sind schätzungsweise 80% der Dyspepsie-Fälle funktioneller Art. Eine nur auf der klinischen Symptomatologie basierende Unterscheidung zwischen organischen und funktionellen Dyspepsien ist jedoch unbefriedigend, weil die Symptome zwar eine hohe Sensitivität, aber nur eine geringe Spezifität für organische Leiden haben.

Patienten mit organisch bedingter Dyspepsie klagen oft über eine zunehmende Einschränkung ihrer körperlichen Fähigkeiten, während Patienten mit funktioneller Dyspepsie häufiger wegen Unwohlsein die täglichen Aktivitäten unterbrechen und öfter einen Arzt konsultieren. Der Leidensdruck der Patienten mit funktioneller Dyspepsie scheint größer zu sein als bei Patienten mit organischen Krankheitsbildern.

Bisher gibt es verschiedene Studien, die sich mit der Kosten-Nutzen-Beziehung unterschiedlicher Management-Strategien der Dyspepsie befassen; jedoch handelt es sich bei diesen Studien in den meisten Fällen um Computersimulationen. Das besondere Interesse in diesen Studien galt der Frage, ob bei sämtlichen neuen Dyspepsiefällen endoskopiert werden sollte oder ob ein differenzierteres Vorgehen sinnvoller ist. Es gibt eine große Zahl prospektiver, kontrollierter Kurzzeitstudien zur Pathophysiologie und Therapie der Dyspepsie. Allerdings fehlen prospektive Langzeitdaten an einer großen Zahl möglichst nicht ausgewählter Patienten mit dyspeptischen Beschwerden. Daher wurde die **pr**ospektive **e**pidemiologische **St**udie der **O**berbauchbeschwerden (PRESTO) konzipiert, die möglichst realitätsnah das gegenwärtige Management und den Verlauf der Dyspepsie dokumentieren sowie die dadurch verursachten Kosten erfassen soll.

Die Autoren von [1] sind der wissenschaftliche Beirat (Vorsitz: Prof. Dr. med. Drs. h. c. M. Classen, Kinikum Rechts der Isar, München) und die Durchführenden der PRESTO-Studie. Die Organisation und Datenerhebung wurde finanziert von der Janssen-Cilag GmbH, Neuss. Die Genehmigung und Begutachtung der Studie erfolgte durch die Ethikkommission der Landesärztekammer Rheinland-Pfalz.

2 Das Design der PRESTO-Studie

Als offene, multizentrische, epidemiologische Datenerhebung durch Hausärzte an Patienten mit Oberbauchbeschwerden ist PRESTO über 2 Jahre konzipiert, mit dem Ziel, den derzeitigen Stand in Bezug auf Klinik, Lebensqualität und Kosten zu erfassen. Die Patienten sollten definierten Kriterien entsprechen [1]. Die Rekrutierungsphase der Studie war das Jahr 1997, und die teilnehmenden

Ärzte wurden gebeten, möglichst konsekutiv geeignete Patienten gemäß den entsprechenden Einschlusskriterien in die Studie aufzunehmen. Der diagnostische und therapeutische Handlungsablauf war den teilnehmenden Ärzten freigestellt.

Bei der Aufnahmevisite wurden demographische Daten, sozioökonomische und persönliche Angaben, Lebensgewohnheiten und Begleiterkrankungen dokumentiert. Bei der Aufnahme sowie den Kontrollvisiten (nach 4 Wochen, 6, 12, 18 und 24 Monaten) notierten die Prüfärzte Symptome, medikamentöse und alternative Therapien, Leistungsfähigkeit und allgemeine Lebensqualität, stressorische Ereignisse und die Befindlichkeitsbeurteilung durch Arzt und Patient. Zur Beurteilung der Symptome, der Lebensqualität und der Befindlichkeitsbeurteilung wurden validierte Messmethoden verwendet. Die Symptome wurden auf einer geringfügig modifizierten DIGEST-Skala bewertet, die Lebensqualität mit dem „Psychological General Well-Being Index“(PGWBI) sowie die allgemeine Befindlichkeitsbeurteilung durch die Visuelle Analog-Skala (VAS), siehe [1].

Auf Grund der erhobenen Daten sollten vor allem drei Haupt- und drei Nebenhypothesen untersucht und geprüft werden.

Haupthypothesen:

1. Die funktionelle Dyspepsie ist eine chronisch-rezidivierende Krankheit.
2. Die medikamentöse(n) Therapie(n) ist (sind) wirksam und ökonomisch.
3. Die symptomatische Klassifikation der Dyspepsie in „Reflux-Typ“, „Ulkus-Typ“, „Dysmotilitäts-Typ“, „Misch-Typ“ ist für die Therapieentscheidung nicht hilfreich.

Nebenhypothesen:

4. Die initiale bildgebende Diagnostik ist der probatorischen Therapie im Hinblick auf das Beschwerdebild und die Kosten nicht überlegen.
5. Alter und psychischer Stress beeinflussen den Verlauf der Beschwerden nicht.
6. Die Ergebnisse dieser Verlaufsstudie sind für die Erstellung oder Anpassung von Leitlinien geeignet.

In der vorliegenden Arbeit wird der Versuch unternommen, Prognosemodelle für Therapieresponder zu erstellen, so dass teilweise Antworten auf die Hypothesen 2., 3. und 5. gegeben werden können. Die Definition eines Therapieresponders sowie die Beschreibung des vorliegenden Datenmaterials erfolgt im nächsten Abschnitt.

3 Datenmaterial und Definition von Therapierespondern

In die PRESTO-Studie wurden 3001 Patienten aufgenommen, von denen noch 2736 zum Therapiekontrollbesuch nach 4 Wochen erschienen. Die Daten dieser 2736 Patienten bilden die Grundlage für die weiteren Analysen.

Ziel dieser Untersuchungen ist es, Prognosemodelle für Therapieresponder zu bestimmen. Auf Grund des vorliegenden Datenmaterials wurden 16 Variablen als mögliche prognostische Variablen identifiziert. Diese 16 Variablen umfassen neben Alter und Geschlecht Angaben über Lebensgewohnheiten (Rauchen, Alkohol, Kaffee, Tee), Vorliegen von Dauerstress bzw. Verarbeiten eines stressorischen Ereignisses, berufliche Tätigkeit und ungünstige Dienstzeiten, Konsultationen wegen Oberbauchbeschwerden in den letzten drei Monaten sowie Anwendung von diagnostischen Verfahren bei diesen Konsultationen, bekannte funktionelle Störungen bzw. Störungen des Essverhaltens, die Typenzuordnung (vgl. Hypothese 3) sowie die initiale Therapieform. Da den Ärzten die Behandlung der Patienten nicht vorgeschrieben war, konnten sie frei über die initiale Therapieform entscheiden. Die gewählten initialen Therapieformen lassen sich drei verschiedenen Gruppen zuordnen: die probatorische, medikamentöse Therapie, Endoskopie mit anschließender medikamentöser Therapie sowie eine „wait-and-see“-Strategie, d. h. es wurde zunächst weder endoskopiert noch medikamentös behandelt.

Die Definition der Therapieresponder orientiert sich an die in Abschnitt 2 erwähnten validierten Messmethoden zur Beurteilung der dyspeptischen Symptome, der Lebensqualität und der Befindlichkeitsbeurteilung. Um die Hypothese 1, dass die funktionelle Dyspepsie eine chronisch-rezidivierende Krankheit ist, in die Definition der Responder mit einzubeziehen, wurden die Daten zu den Symptomen, der Lebensqualität und der Befindlichkeitsbeurteilung bis zum vierten Untersuchungszeitpunkt, d. h. bis ein Jahr nach Einschluss in die Studie, betrachtet. Fehlende Werte bei einem Untersuchungszeitpunkt wurden durch den Wert der vorherigen Untersuchung ersetzt. Dies entspricht dem Prinzip des „last observation carried forward“[2], welches standardmäßig in klinischen Versuchen verwendet wird.

Es werden im Folgenden zwei Typen von Respondern betrachtet, zum einen Responder, die durch die Reduktion der dyspeptischen Beschwerden definiert werden, und zum anderen Responder, die durch die Verbesserung der Lebensqualität gekennzeichnet sind. Zur Beurteilung der dyspeptischen Beschwerden stehen drei Variablen zur Verfügung: die Symptomsumme der modifizierten DIGEST-Skala, eine 5-Punkte-Skala der Beurteilung der Beschwerden durch den Arzt sowie die Beurteilung der Beschwerden durch den Patienten anhand der Visuellen Analog-Skala. Die Responder bezüglich der Reduktion der Beschwerden sind so definiert, dass in mindestens zwei der drei Kriterien eine

Verbesserung bezogen auf den Untersuchungzeitraum von einem Jahr eingetreten ist und dass keine Verschlechterung in Bezug auf die Anfangswerte bei den weiteren Untersuchungen aufgetreten ist. Von den 2736 Patienten sind auf Grund dieser Definition 1646 Patienten (60.2%) als Responder bezüglich der Beschwerdenreduktion identifiziert worden.

Die Beurteilung der Lebensqualität erfolgt anhand des „Psychological General Well-Being Index"(PGWBI), der sich aus 22 Antworten bezüglich ängstlichkeit, Depressivität, Gesundheit, Selbstkontrolle, Vitalität und Wohlbefinden berechnet. Das zu analysierende Kollektiv reduziert sich auf Grund des Nichtausfüllens des Fragebogens auf 2032 Patienten. Als Responder bezüglich der Lebensqualität werden die Patienten bezeichnet, die sich mindestens einmal im PGWBI verbessert haben und sich nicht gegenüber ihrem Ausgangswert verschlechtert haben. Mit dieser Definition werden 1053 Patienten (51.8%) als Responder bezüglich der Verbesserung der Lebensqualität identifiziert. Von diesen 1053 Patienten sind 721 Patienten auch Responder bezüglich der Beschwerdenreduktion.

4 Prognosemodelle für Therapieresponder

Mit Hilfe des SAS Enterprise Miners [3] sind Prognosemodelle für die Therapieresponder erstellt worden. Auf Grund der „geringen"Anzahl von Patienten ist auf eine Aufteilung der Datensätze in Trainings- und Validierungsdatensätze verzichtet worden, statt dessen sind jeweils die gesamten Datensätze trainiert worden. Der einzige intervallskalierte prognostische Faktor, das Alter, ist im *Transform Variable*-Knoten in vier Altersklassen aufgeteilt worden, wobei die Grenzen der Altersklassen durch das untere Quartil, den Median und das obere Quartil der empirischen Altersverteilung festgelegt wurden. Die fehlenden Werte in den binären bzw. nominalskalierten prognostischen Faktoren wurden im *Replacement*-Knoten mit der Methode *Tree Imputation* jeweils unter Berücksichtigung aller anderen prognostischer Faktoren ersetzt. Die Anzahl der fehlenden Werte in einem prognostischen Faktor betrug höchstens 6%.

Zur Erstellung der Prognosemodelle wurden die Data Mining Verfahren Logistische Regression, Entscheidungbaumverfahren und Künstliche Neuronale Netze herangezogen [4]. Zunächst wurden Prognosemodelle mit Hilfe der Logistischen Regression und von Entscheidungsbaumverfahren berechnet. Neben der Prognosegüte war dabei auch die Reduktion auf wesentliche prognostische Faktoren von Interesse.

Bei der Logistischen Regression wurde stets die schrittweise Prozedur zur Variablenselektion benutzt, wobei als Selektionskriterium die Minimierung des *Cross Validation Error* gewählt wurde. Im Knoten für die Entscheidungsbaumverfahren wurden die Einstellungen so gewählt, dass von jedem Knoten genau

zwei Äste abzweigen und die Trennungen auf Grund des χ^2-Tests mit einem Signifikanzniveau von 0.1 bestimmt werden. Durch diese Vorgaben wird erreicht, dass der Baum nicht zu tief wird, so dass das sogenannte „overfitting“ beim Trainieren der entsprechenden Datensätzen unterbunden wird. Zusätzlich wurde die maximale Tiefe der Entscheidungsbäume auf vier festgelegt.

Schließlich wurden mit den als wesentlich identifizierten prognostischen Faktoren aus der Logistischen Regression und dem Entscheidungsbaumverfahren noch Künstliche Neuronale Netze trainiert. Als Architektur dieser Netze wurde stets ein Multilayer Perceptron (MLP) mit einer verdeckten Schicht verwendet, wobei die Anzahl der versteckten Neuronen in der verdeckten Schicht variiert wurde. Mit Hilfe der MLPs sollte überprüft werden, ob sich die Prognosegüte der schrittweisen Logistischen Regression und des Entscheidungsbaumverfahrens noch wesentlich verbessern lässt. Die nächsten Abschnitte enthalten die Ergebnisse der Data Mining Verfahren für die beiden Respondertypen.

4.1 Reduktion der Beschwerden

Auf Grund der Logistischen Regression mit schrittweiser Variablenselektion wurden sechs wesentliche prognostische Faktoren identifiziert: Initiale Therapieform, Tee trinken, Konsultationen wegen Oberbauchbeschwerden in den letzten drei Monaten, Anwendung diagnostischer Verfahren bei Konsultationen in den letzten drei Monaten, ungünstige Dienstzeiten sowie berufliche Tätigkeit (8 Ausprägungen). Der wichtigste prognostische Faktor ist dabei die initiale Therapieform, wobei die probatorische, medikamentöse Therapie und vor allem die Endoskopie mit anschließender medikamentöser Therapie der „wait-and-see“-Strategie signifikant überlegen sind. Weiterhin ist es günstig, kein Tee-Trinker zu sein und keinen Arzt wegen Oberbauchbeschwerden in den letzten drei Monaten konsultiert zu haben, um eine gute Responderprognose zu erzielen. Mit Hilfe dieses Prognosemodells erreicht man eine hohe Sensitivität für Responder bezüglich der Reduktion der dyspeptischen Beschwerden, jedoch eine schlechte Spezifität.

Bei der Berechnung des Entscheidungsbaumes für die Responder bezüglich der Beschwerdereduktion zeigte sich, dass die erste binäre Trennung des Patientenkollektivs anhand der initialen Therapieform erfolgt. Patienten, die zunächst die „wait-and-see“-Strategie verordnet bekamen, hatten deutlich schlechtere Responderraten (48.5%) als Patienten, die einer der anderen beiden initialen Therapieformen bekamen. Von den Patienten mit der „wait-and-see“-Strategie hatten Patienten, die älter als 61 Jahre waren, mit 28.8% die schlechteste Responderrate. Bei den Patienten, die medikamentös behandelt oder endoskopiert wurden, lag die Responderrate bei Patienten ohne Konsultationen wegen Oberbauchbeschwerden in den letzten 3 Monaten bei 68%; waren die Patienten

auch noch Raucher, so erhöhte sich die Responderrate auf 74.2%. Alle weiteren Blätter im Entscheidungsbaum brachten keine wesentlichen Erkenntnisse. Mit Hilfe des Entscheidungsbaums konnten somit für einige Subgruppen gute Prognosewerte extrahiert werden. Wie bei der Logistischen Regression ist auch beim Entscheidungsbaum die initiale Therapieform der wichtigste prognostische Faktor.

Mit dem *Neural Network*-Knoten im SAS Enterprise Miner wurden MLPs für die Responder bezüglich der Beschwerdenreduktion trainiert. Ein MLP hatte als Input-Variablen die Variablen aus der Logistischen Regression mit schrittweiser Variablenselektion. Die verdeckte Schicht in dieser Architektur enthielt drei versteckten Neuronen. Ein zweites MLP hatte als Input-Variablen die Variablen aus dem Entscheidungsbaum: Initiale Therapieform, Alter (vier Altersklassen), Konsultationen wegen Oberbauchbeschwerden in den letzten drei Monaten, Rauchen, Dauerstress und ungünstige Dienstzeiten. Die Anzahl der versteckten Neuronen wurde in dieser Architektur mit sechs gewählt.

Die Abbildung 1 enthält nun die kumulativen Response Lift-Charts für die vier Prognosemodelle.

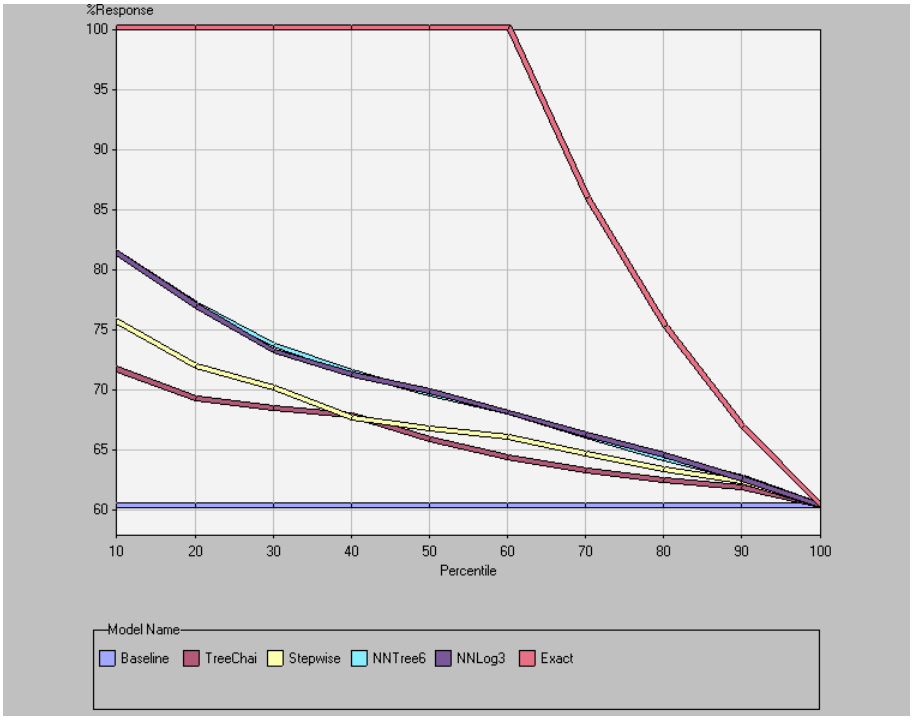

Abbildung 1: Kumulative Response Lift-Charts von Prognosemodellen für Responder bzgl. Beschwerdenreduktion, TreeChai = Entscheidungsbaum, Stepwise = Logistische Regression, NNTree6 = MLP mit 6 versteckten Neuronen, Variablen aus Entscheidungsbaum, NNLog3 = MLP mit 3 versteckten Neuronen, Variablen aus Logistischer Regression

Wie der Abbildung 1 zu entnehmen ist, besitzt die Logistische Regression eine höhere Sensitivität für Responder bezüglich der Reduktion des dyspeptischen Beschwerden als der Entscheidungsbaum. Mit den beiden MLPs lässt sich die Prognosegüte noch verbessern, wobei jedoch das MLP mit den selektierten Variablen aus dem Entscheidungsbaum doppelt so viele versteckte Neuronen benötigt wie das MLP mit den selektierten Variablen aus der Logistischen Regression, um die gleiche Prognosequalität zu erhalten.

4.2 Verbesserung der Lebensqualität

Die Logistische Regression mit schrittweiser Variablenselektion lieferte zwei wesentliche prognostische Faktoren für Responder bezüglich der Verbesserung der Lebensqualität: Verarbeiten eines stressorischen Ereignisses und berufliche Tätigkeit. Das Verarbeiten eines stressorischen Ereignisses wie z. B. Tod oder Erkrankung eines Verwandten oder Freundes, Verlust der Arbeit, Scheidung etc. hat den größten prognostischen Einfluss darauf, ob sich die Lebensqualität des Patienten verbessert.

Im Entscheidungsbaum war ebenfalls das Verarbeiten eines stressorischen Ereignisses der wichtigste prognostische Faktor, denn anhand dieser Variable wurde der erste binäre Split durchgeführt. In der Gruppe der Patienten mit stressorischem Ereignis hatten die Patienten, die jünger als 61 Jahre waren, mit 58.7% die höchste Responderrate; diese bildeten mit 805 Patienten auch noch die größte Patientengruppe in einem Blatt des Entscheidungsbaums. Bei Patienten ohne stressorisches Ereignis erreichte die Gruppe von Patienten, die keinen Tee tranken und keinen Arzt wegen Oberbauchbeschwerden in den letzten drei Monaten konsultiert hatten, mit 56.8% die höchste Responderrate.

Analog zum vorherigen Abschnitt lassen sich die Prognosemodelle noch mit Hilfe von Künstlichen Neuronalen Netzen verbessern. Auf eine Darstellung wird hier jedoch verzichtet. In der Abbildung 2 sind daher nur die Lift-Charts für die Modelle aus der Logistischen Regression und aus dem Entscheidungsbaumverfahren dargestellt.

Anhand der Abbildung 2 erkennt man, dass wie im vorherigen Abschnitt die Logistische Regression eine größere Sensititvität für Responder besitzt als das Entscheidungsbaumverfahren. Die im Entscheidungsbaum gefundenen Subgruppen mit den höchsten Responderraten sind relativ gross und diese höchsten Responderraten zu klein, um in den ersten Dezilen mit der Logistischen Regression basierend auf nur zwei prognostische Faktoren konkurrieren zu können.

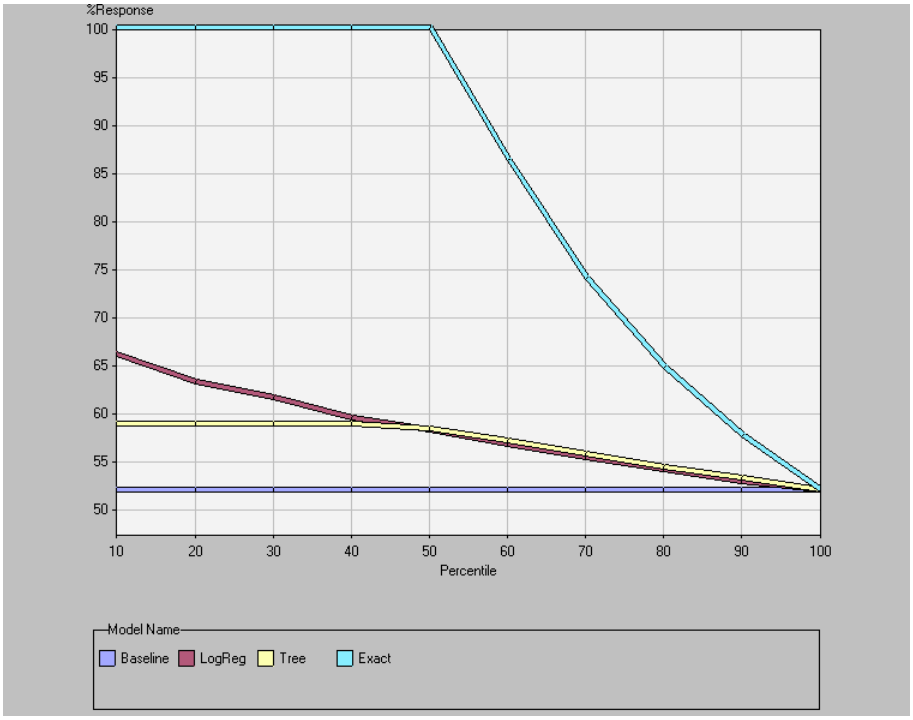

Abbildung 2: Kumulative Response Lift-Charts von Prognosemodellen für Responder bzgl. Verbesserung der Lebensqualität, LogReg = Logistische Regression, Tree = Entscheidungsbaumverfahren

5 Diskussion

Mit Hilfe der Data Mining Methoden im SAS Enterprise Miner lassen sich Antworten für einige Hypothesen der PRESTO-Studie ableiten. Bezüglich der Reduktion der Beschwerden ist die initiale Therapieform der wichtigste prognostische Faktor. Dies ist sowohl das Ergebnis der Logistischen Regression als auch des Entscheidungsbaumverfahrens. Die initiale probatorische, medikamentöse Therapie erweist sich als wirksame Therapie, auch wenn die Endoskopie mit anschließender medikamentöser Therapie noch größere Responderraten hervorbringt. Den ökonomischen Aspekt können wir an dieser Stelle nicht beurteilen.

Die symptomatische Klassifikation der Dyspepsie spielte in keinem der Prognosemodelle eine Rolle, so dass die Gültigkeit der Hypothese 3 mit den vorliegenden Analysen indirekt untermauert wird. Ebenfalls sind Alter und psychischer Stress keine wesentlichen prognostischen Faktoren in Bezug auf die Reduktion der Beschwerden.

Für die Verbesserung der Lebensqualität, bewertet mit dem PGWBI, ist das Verarbeiten eines stressorischen Ereignisses der wichtigste prognostische Faktor. Ein solches Einzelereignis führt augenscheinlich zu einer kurzfristigen Ver-

minderung der Lebensqualität. Man muss jedoch insgesamt beachten, dass nur etwa 50% der Patienten als Responder bezüglich der Verbesserung der Lebensqualität eingestuft wurden, so dass die funktionelle Dyspepsie einen gewissen Leidensdruck auf die Patienten ausübt.

Literatur

[1] Allescher, H.D., Adler, G., Hartung, J., Manns, M.P., Riemann, J.F., Wienbeck, M., Classen, M. und die PRESTO-Studiengruppe (1999). Prospektive Epidemiologische Studie der Oberbauchbeschwerden (PRESTO). Grundlagen und erste Ergebnisse. Deutsche Medizinische Wochenschrift 124, 443-450.

[2] Committee for Proprietary Medicinal Products (2001). Points to Consider on Missing Data. The European Agency for the Evaluation of Medicinal Products, London, CPMP/EWP/1776/99.

[3] SAS Enterprise Miner^{TM}, Version 4.0. SAS Institute Inc., Cary, NC.

[4] Hartung, J., Knapp, G. (2001). Data Mining. In: Power Tools. Management-, Beratungs- und Controllinginstrumente, Hrsg. D. Schneider und P. Pflaumer, Gabler, Wiesbaden, 189-201.