

BARMER GEK

Gesundheitswesen aktuell 2010

Beiträge und Analysen

herausgegeben von Uwe Repschläger,
Claudia Schulte und Nicole Osterkamp



Impressum

Das Werk einschließlich aller seiner Teile ist Eigentum der BARMER GEK. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung der BARMER GEK unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Im Sinne der besseren Lesbarkeit wurde überwiegend die grammatikalisch männliche Sprachform gewählt. Wenn im Text die männliche Sprachform genannt ist, ist damit sowohl die männliche als auch die weibliche Sprachform gemeint.

BARMER GEK Gesundheitswesen aktuell 2010

© 2010 BARMER GEK

herausgegeben von

Uwe Repschläger, Claudia Schulte und Nicole Osterkamp

ISBN 978-3-9812534-2-9

Realisation und Design: 37 Grad GmbH, Düsseldorf

Redaktion: 37 Grad GmbH, Düsseldorf

produziert und gedruckt in Deutschland

Alle Rechte vorbehalten

Änderungen und Irrtümer vorbehalten

Inhaltsverzeichnis

Vorwort	8
Editorial.....	10

I. Systemfragen und Wettbewerb

Preisregulierung von patentgeschützten Arzneimitteln. Welchen Beitrag kann die Kosten-Nutzen-Bewertung leisten?	16
Mathias Kifmann	
QALYs in der Kosten-Nutzen-Bewertung. Rechnen in drei Dimensionen.....	32
Klaus Koch, Andreas Gerber	
Die Erstellung der Ambulanten Kodierrichtlinien	50
Ingolf Berger	
Vom Zusatzbeitrag zur Gesundheitspauschale. Das niederländische Gesundheitssystem als Vorbild?	70
Frank Schulze Ehring	
Frühwarnsysteme in der Gesetzlichen Krankenversicherung. Die Rahmenbedingungen	90
Falk König, Patrick Florath, Uwe Repschläger, Frank Romeike	
Zur Rolle und Funktion der alten Spitzenverbände. Perspektive für eine zukunftsfähige Verbändelandschaft in einem wettbewerbsorientierten GKV-System	106
Uwe Repschläger	

Die Abbildung von Multimorbidität im morbiditätsorientierten Risikostrukturausgleich	126
Claudia Schulte	

II. Gestaltung der Versorgung

Kollektivverträge und selektive Vollversorgungsverträge in der ambulanten ärztlichen Versorgung. Die Frage nach dem „Entweder-oder“	148
Andreas Köhler	

Bereinigung der vertragsärztlichen Vergütung. Stand der Verfahrensentwicklung und Relevanz für einen fairen Wettbewerb zwischen kollektiv- und selektivvertraglicher Versorgung	160
Antje Schwinger, Hans-Dieter Nolting	

Achtzehn Arztkontakte im Jahr. Hintergründe und Details.....	176
Boris von Maydell, Thilo Kosack, Uwe Repschläger, Christoph Sievers, Rebecca Zeljar	

Der deutsche Arzneimittelmarkt. Reform- oder neuordnungsbedürftig?	192
Detlef Böhler, Karl-Heinz Neumann	

Das Fibromyalgiesyndrom. Dilemma zwischen Leitlinie und Versorgungsrealität	212
Ursula Marschall, Andreas Wolik	

Adipositaschirurgie – notwendige Therapie oder Lifestyle? Spannungsfeld zwischen Medizin und Krankenkasse	240
Hans Hauner, Ursula Marschall, Michael Lex, Andreas Wolik, Henrik Schwandrau	

Berechnung von Hospitalisierungswahrscheinlichkeiten. Die Methode der Zufallswälder und der Vergleich mit gängigen statistischen Klassifikationsverfahren	266
Thilo Kosack, Andreas Wolik	

III. Anhang

Abkürzungsverzeichnis	290
Autorenverzeichnis	292

Thilo Kosack, Andreas Wolik

Berechnung von Hospitalisierungswahrscheinlichkeiten

Die Methode der Zufallswälder und der Vergleich mit gängigen statistischen Klassifikationsverfahren

Den Ausgangspunkt für die Darstellung und Anwendung der noch relativ neuen Methode der Zufallswälder bildet in diesem Beitrag die passgenaue Gestaltung des Versorgungsmanagements von Krankenkassen für eine ressourcenoptimale Verwendung von Mitteln. Die Methodik soll bessere und stabilere Prognosen im Hinblick auf die Versichertenversorgung ermöglichen und stellt eine Erweiterung des Entscheidungsbaumverfahrens dar. Der Beitrag analysiert und vergleicht die prognostische Güte verschiedener Analysetechniken (Entscheidungsbäume, logistische Regressionen und Zufallswälder). Darüber hinaus werden der Nutzen einer Voraggregation der Daten durch eine Grouper-Software bei den verschiedenen Methoden analysiert sowie Vor- und Nachteile dieser Vorgehensweisen präsentiert.

Einleitung

Die möglichst genaue Prognose von zukünftig besonders teuren und kostenintensiven Fällen und die Einbindung dieser Versicherten in ein Versorgungsmanagement mit geeigneten Maßnahmen sind wichtige Ziele zur Kostenminderung. Versichertenbezogene Kostensteigerungen ergeben sich insbesondere aufgrund von ungünstigen oder krisenhaften Krankheitsverläufen und den damit oftmals einhergehenden Krankenhauseinweisungen (den sogenannten Hospitalisierungen). Hierbei ist ein Risikowert, der die Eintrittswahrscheinlichkeit einer Hospitalisierung wiedergibt, ein geeignetes Mittel, um Ressourcen gezielt und effizient einzusetzen, da eine Konzentration auf die wahrscheinlichsten Eintrittsfälle möglich ist. Somit können die richtigen Versicherten angesprochen werden und die oben beschriebenen ungünstigen Verläufe durch geeignete Maßnahmen vermieden werden.

Die Versicherteninformationen aus den Abrechnungsdaten sind als Informationsbasis zur Risikoprognose besonders geeignet. Sie liegen in einem sehr hohen Detaillierungsgrad auf der Diagnose- und der Arzneimittelverordnungsebene vor. Bisher war es notwendig, diese Informationen mittels einer Grouper-Software zu aggregieren und damit eine Informationsbasis zu erzeugen, die für gängige Data-Mining-Methoden, wie beispielsweise die Entscheidungsbäume, geeignet ist (siehe Wolik 2008).

Die Suche nach immer besseren und stabileren Prognosen ist besonders im Fall von Hochkostenrisiken von Relevanz, da selbst eine geringe Steigerung in der Prognosegüte einen deutlich positiven Kosteneffekt bewirken kann. Daher wird im Folgenden ein bisher wenig bekannter Ansatz der Prognosemodellbildung vorgestellt: die Zufallswälder.

Diese Methode ist eine Weiterentwicklung eines Entscheidungsbaumverfahrens (Classification And Regression Tree – CART) und besitzt den konzeptionellen Vorteil, sich auch für Daten mit sehr hoher Informationstiefe wie Diagnose- und Arzneimittelinformationen zu eignen. Hierdurch bietet sich die Möglichkeit, Prognosemodelle mit hoher Güte auch ohne die Zuhilfenahme einer Grouper-Software zu erstellen. Dies ist von besonderem Interesse, da Informationen bei der Aggregation mithilfe von Grouping-Techniken verloren gehen und Zusammenhänge auf detaillierter Diagnose- und Arzneimittelenebene eine bessere Prognose ergeben können. Hierzu werden die klassischen Methoden des Entscheidungsbaumverfahrens und der logistischen Regression mit dem Verfahren der Zufallswälder verglichen und überprüft, in welchen Fällen die Nutzung einer Grouper-Software notwendig ist. Darüber hinaus wird untersucht, welchen Effekt statistische Methoden zur Variablenselektion in Ergänzung zu den genutzten Methoden haben.

Datengrundlage

Die Datenbasis für die empirische Analyse in diesem Artikel stellen die Daten der BARMER von an Diabetes mellitus leidenden Versicherten in

den Jahren 2005 und 2006 dar. Diese umfassen sämtliche sowohl im ambulanten wie auch im stationären Bereich gestellten Diagnosen in Form von ICD-Codes sowie die verordneten Wirkstoffe samt Dosis und Verordnungsdauer. Darüber hinaus liegen noch Informationen vor, ob obige Versicherte im Jahr 2007 hospitalisiert wurden. Zudem wurden bestimmte Multimorbiditäten durch das gemeinsame Vorliegen bestimmter Diagnosen und oder Arzneimittelgaben definiert, die als weitere Parameter vorgegeben sind.

Funktionsweise von Zufallswäldern

Als Methoden zur Berechnung der Hospitalisierungswahrscheinlichkeit werden im Folgenden Entscheidungsbäume mittels CART (Breiman et al. 1984), logistische Regression (Hosmer und Lemeshow 1989) und Zufallswälder (Breiman 2001) verwendet. An dieser Stelle wird kurz die Funktionsweise von Zufallswäldern erläutert, da diese methodisch ungewöhnlich und darüber hinaus noch sehr wenig bekannt sind.

Zufallswälder basieren auf Entscheidungsbäumen mittels CART. Ein Zufallswald besteht aus einer größeren Menge von Entscheidungsbäumen, deren Anzahl von der Rechenkapazität und der Variablenzahl abhängt, jedoch laut Empfehlungen in Fachliteratur wie beispielsweise Berk (2008) nicht unter 500 liegen sollte. Grundsätzlich gilt, dass ein Zufallswald von einer wachsenden Anzahl von Entscheidungsbäumen nur profitieren kann. Demnach sollten so viele Bäume erzeugt werden, wie es die technischen und zeitlichen Restriktionen erlauben. Die Wahrscheinlichkeit für das Eintreten eines Ereignisses ergibt sich bei einem Zufallswald aus den Ergebnissen der einzelnen Entscheidungsbäume. Wenn beispielsweise jeder zweite Baum als Ergebnis den Eintritt des interessierenden Ereignisses voraussagt, so liegt seine durch den Wald vorhergesagte Prognosewahrscheinlichkeit bei 50 Prozent.

Während der Erzeugung einzelner Entscheidungsbäume eines Zufallswaldes finden an zwei Stellen Zufallsauswahlen statt. Zunächst wird eine zufällige Auswahl an Datensätzen (Bagging) verwendet. Es werden

zum Erstellen eines einzelnen Baumes also nicht alle der möglichen Daten verwendet, sondern lediglich eine Zufallsstichprobe (Ziehen mit Zurücklegen). Aus dem für den jeweiligen Baum zufällig erzeugten Datensatz findet anschließend an mehreren Stellen eine Zufallsauswahl der Variablen statt, sodass zur Generierung der einzelnen Bäume unter Umständen verschiedene Variablen genutzt werden. Diese Zufallsauswahlen sorgen erst dafür, dass nicht alle erzeugten Bäume identisch sind. Empirische Untersuchungen zeigen, dass durch diese Vorgehensweise insbesondere die Variantenhäufigkeit der in den Bäumen genutzten Variablen gesteigert wird und für die hervorragende Vorhersagegüte von Zufallswäldern verantwortlich ist (siehe Berk 2008).

Ein weiterer Vorteil von Zufallswäldern zeigt sich darin, dass auch bei einer geringen Anzahl von Untersuchungseinheiten (Patienten, Versicherte) stabile und gute Prognosen generierbar sind. Die Vorteile der Zufallswälder und ihre besonders gute Anwendbarkeit bei Diagnoseinformationen haben daher in den letzten Jahren zu einer Vielzahl von Anwendungen im medizinischen Kontext geführt (siehe Shi et al. 2005, Wu, Wu und Li 2008 sowie Yang et al. 2009).

Methodisches Vorgehen

Vor der Berechnung der Hospitalisierungswahrscheinlichkeiten werden zunächst zwei Stichproben von Versicherten gezogen. Bei der ersten Stichprobe wird die Grouper-Software der Firma DxCG verwendet, um die Zahl der bei den Versicherten beobachteten Variablen zu verringern. Bei der zweiten Stichprobe wird auf den Einsatz der Grouper-Software verzichtet. Hinsichtlich der Stichprobengröße gibt es abhängig von der Anzahl der vorhandenen Variablen gewisse Restriktionen: Die Ursache dafür liegt in der verwendeten statistischen Analysesoftware „R“ in der Version 2.6.0. Es kann lediglich eine begrenzte Datenmenge verarbeitet werden, sodass die erste Stichprobe einen Umfang von 50.000 Versicherten und 427 erklärenden Variablen hat, während die zweite Stichprobe einen Umfang von 15.000 Versicherten und 13.450 erklärenden Variablen besitzt.

Wegen der Neuartigkeit insbesondere der Methode der Zufallswälder kann jedoch keine leistungsfähigere Software verwendet werden. Die beiden Stichprobenumfänge werden unterschiedlich gewählt, da selbst bei Verwendung von leistungsfähigerer Software nicht zu erwarten ist, dass ein Datensatz mit 13.450 Variablen bei den untersuchten Methoden und einem maximal möglichen Umfang von etwa sieben Millionen Zeilen (eine pro Versicherten) den gleichen Umfang besitzen kann wie ein Datensatz mit lediglich 427 Variablen. Aus diesem Grund werden die Stichprobenumfänge jeweils größtmöglich gewählt, um eine größere Datenbasis zu erzielen. Die beiden Stichproben werden selbst wiederum aufgeteilt in jeweils einen Lerndatensatz, der zwei Drittel der Daten umfasst, und einen Testdatensatz, der aus einem Drittel der Daten besteht.

Bei beiden Stichproben werden neben den vollständigen Modellen mit allen Variablen auch noch Modelle mit weniger als der maximalen Anzahl von Variablen getestet. Im Fall der zweiten Stichprobe ist es sogar zwangsläufig notwendig, ein reduziertes Modell zu nutzen. Dazu werden verschiedene statistische Methoden zur Reduktion der Anzahl der Variablen verwendet, wobei wegen der durch die Software vorgegebenen Restriktionen eine Einschränkung auf die folgenden sehr einfachen Methoden stattfinden muss. Zur Modellreduktion werden folgende Techniken genutzt:

- Die Auswahl der einzelnen erklärenden Variablen erfolgt abhängig von ihrer Korrelation mit der Zielvariablen. Dabei werden alle Variablen beibehalten, deren Korrelation mit der Zielvariablen mindestens dem Y -fachen der durchschnittlichen Korrelation ($Y \times \bar{\text{Korrelation}}$) entspricht. Dabei kann die Zahl der Variablen über den Zahlenwert von Y gesteuert werden. Je größer dieser ist, desto weniger Variablen sind im Modell enthalten.
- Die Ergebnisse einer Klassifikation mit je nur einer erklärenden Variablen. Dabei werden diejenigen Variablen beibehalten, bei denen eine Klassifikation ein besseres Ergebnis als eine Zufallsauswahl erbringt.

- Algorithmen, welche in den drei verwendeten Methoden zur Berechnung der Wahrscheinlichkeit bereits implementiert sind. Für jede der Methoden Entscheidungsbäume, Zufallswälder und logistische Regression gibt es dabei eine jeweils andere Möglichkeit. Diese Algorithmen berechnen auf zum Teil sehr aufwendige Weise Zahlenwerte, die den Beitrag einer Variablen zur Klassifikation angeben. Es können dann diejenigen Variablen ausgewählt werden, die einen bestimmten Mindestzahlenwert besitzen.

Für nähere Informationen zu den ersten beiden Auswahlverfahren sei auf Guyon und Elisseeff (2003) verwiesen. Die Funktion der implementierten Methoden findet sich in der Literatur zur jeweiligen Methode, das heißt im Fall von Entscheidungsbäumen in Breiman et al. (1984), bei Zufallswäldern in Breiman (2001) und für logistische Regression in Hosmer und Lemeshow (1989). Zur Messung der Qualität der Vorhersage wird die Trefferquote der korrekt vorhergesagten Hospitalisierungen verwendet. Die Trefferquote gibt den prozentualen Anteil der korrekt vorhergesagten Hospitalisierungen an den insgesamt von den Methoden vorhergesagten Hospitalisierungen an.

Bei der Differenz der Trefferquoten muss berücksichtigt werden, dass ein Unterschied von einem Prozentpunkt bei einer Vorhersage für 10.000 Personen bereits eine korrekte oder nicht korrekte Vorhersage von 100 Personen bedeutet. Insoweit sind auf den ersten Blick gering erscheinende Differenzen zwischen den Trefferquoten von großer Bedeutung. Ausgehend von diesem Zahlenbeispiel von 100 Personen und einer angenommenen Erfolgsquote von 30 Prozent für die angewandten Maßnahmen, könnte für 30 Personen eine kostspielige Hospitalisierung eingespart werden.

Eine Hospitalisierung wird für die Versicherten mit der höchsten Hospitalisierungswahrscheinlichkeit vorhergesagt, und zwar für so viele Versicherte, wie tatsächliche Hospitalisierungen in dem Datensatz ablesbar sind. Zum Vergleich der Ergebnisse auf einer Stichprobe wird

darüber hinaus ein statistischer Signifikanztest (siehe Dietterich 1997) verwendet, der die Ergebnisse von je zwei Modellen paarweise vergleicht.

Der Vergleich der drei Methoden läuft in mehreren Stufen ab. Zunächst wird für jede der drei Methoden auf der jeweiligen Stichprobe das beste Modell identifiziert, also dasjenige Modell, auf welchem die Methode die besten Ergebnisse liefert. Anschließend werden die Ergebnisse der drei Methoden miteinander verglichen. Es wird dabei für jede Methode das beste zuvor erzeugte Ergebnis verwendet. Zuletzt werden noch die Ergebnisse der drei Methoden auf den beiden Stichproben verglichen, wiederum unter Berücksichtigung der besten Ergebnisse.

Erste Stichprobe: Entscheidungsbäume

Betrachtet man die Ergebnisse einer Vorhersage der Hospitalisierungswahrscheinlichkeit auf der ersten Stichprobe unter Verwendung des Entscheidungsbaum-Verfahrens nach dem CART-Algorithmus, so zeigen sich die in Tabelle 1 dargestellten Trefferquoten bei den vier verschiedenen zugrunde gelegten Modellen.

Tabelle 1: Trefferquoten der verschiedenen Entscheidungsbaummodelle auf der ersten Stichprobe

Modellreduktion	Anzahl Variablen	Trefferquote (in Prozent)
volles Modell	427	37,58
zweifache durchschnittliche Korrelation	56	37,42
Einzelklassifikation	27	38,79
implementierter Algorithmus	90	37,58

Quelle: eigene Berechnung

Die beste Trefferquote in Höhe von 38,79 Prozent weist das Modell mit nur 27 erklärenden Variablen auf, während das Modell mit 56 Variablen mit 37,42 Prozent die schlechteste Trefferquote besitzt. Das volle Modell und das Modell mit 90 Variablen besitzen jeweils eine dazwischenliegende Quote von 37,58 Prozent. Eine Überprüfung auf Signifikanz hinsichtlich der Trefferquoten fällt mit Ausnahme des Vergleichs der beiden Modelle mit identischer Trefferquote positiv aus. Für die erste Stichprobe, die auf der Technik von Entscheidungsbäumen basiert, ist das Modell mit 27 Variablen, welche mittels Einzelklassifikation ausgewählt wurden, besser als die übrigen Modelle.

Erste Stichprobe: Zufallswälder

Bei der Betrachtung von Zufallswäldern muss die Anzahl der Versicherten des Lerndatensatzes nochmalig reduziert werden. Um die volle Anzahl von 427 Variablen zu berücksichtigen, wird der Umfang des Lerndatensatzes auf 20.000 Beobachtungen reduziert. Zusätzlich wird geprüft, ob eine Steigerung der Anzahl der Bäume über die standardmäßigen 500 hinaus trotz der zwangsläufig damit verbundenen weiteren Verringerung der Anzahl der Beobachtungen bessere Ergebnisse erbringt. Letzteres bestätigt sich jedoch nicht.

In die weitere Betrachtung werden also die Ergebnisse von vier Zufallswäldern einbezogen. Einer mit dem vollen Modell aus 427 Variablen bei 20.000 Beobachtungen im Lerndatensatz und auf der anderen Seite drei Zufallswälder, welche auf Modellen mit einer vorangegangenen Variablenreduktion und dem vollständigen Lerndatensatz basieren. Die Ergebnisse dazu zeigt Tabelle 2.

Tabelle 2: Trefferquoten der Zufallswälder auf der ersten Stichprobe nach Variablenreduktion

Modellreduktion	Anzahl Variablen	Trefferquote (in Prozent)
volles Modell	427	39,72
einfache durchschnittliche Korrelation	150	40,31
eineinhalbfache durchschnittliche Korrelation	96	40,31
implementierter Algorithmus	100	39,43

Quelle: eigene Berechnung

Die Variablenreduktion ergibt sich ausschließlich durch die Auswahl gemäß der Korrelation und der zu den Zufallswäldern gehörenden implementierten Algorithmen. Wegen der Zufallsauswahl von Variablen ist es bei Zufallswäldern nicht sinnvoll, lediglich eine erklärende Variable zu verwenden. Daher wird bei der Korrelation die einfache beziehungsweise eineinhalbfache durchschnittliche Korrelation verwendet, da eine weitere Reduzierung der Variablenanzahl bei Zufallswäldern keine besseren Ergebnisse zeigt. Dies folgt insbesondere aus dem postulierten Vorteil von Zufallswäldern, sehr gut mit einer großen Variablenanzahl umgehen zu können.

Die Zufallswälder, welche nach einer Auswahl gemäß Korrelation erstellt wurden, weisen mit 40,31 Prozent eine fast einen Prozentpunkt höhere Trefferquote auf als der Zufallswald, bei dem die Variablen mittels des implementierten Algorithmus ausgewählt wurden, dessen Trefferquote 39,43 Prozent beträgt. Die bestehenden Unterschiede werden als statistisch signifikant ausgewiesen. Dagegen wird der Unterschied in der Trefferquote zwischen den beiden mittels Korrelation reduzierten Zufallswäldern und dem auf dem vollen Modell basierenden Zufallswald trotz der Differenz von mehr als 0,5 Prozentpunkten nicht als signifikant gekennzeichnet. Es kann also statistisch nicht sicher gesagt werden, dass die Zufallswälder basierend auf dem Modell mit

Variablenreduktion besser sind als diese, auch wenn sie die deutlich beste Trefferquote aufweisen.

Erste Stichprobe: logistische Regression

Bei der logistischen Regression werden lediglich zwei Modelle erzeugt, da sie einem Vergleich der moderneren Methoden der Entscheidungsbäume und der Zufallswälder mit einer Standardmethode dient. Die logistische Regression wird einmal auf dem vollen Modell mit allen Variablen durchgeführt und ein weiteres Mal auf einem Modell, welches nur noch die Variablen mit signifikantem Einfluss beinhaltet. An den in Tabelle 3 dargestellten Ergebnissen lässt sich ablesen, dass die Regression auf dem reduzierten Modell mit 40,24 Prozent Trefferquote etwa einen Prozentpunkt besser ist als die Regression auf dem vollen Modell. Dieser deutliche Unterschied wird auch als statistisch signifikant gekennzeichnet.

Tabelle 3: Trefferquote der logistischen Regression auf der ersten Stichprobe

Modellreduktion	Anzahl Variablen	Trefferquote (in Prozent)
volles Modell	427	39,26
implementierter Algorithmus	52	40,24

Quelle: eigene Berechnung

Erste Stichprobe: Vergleich der drei Methoden

Abschließend werden die drei zur Vorhersage der Hospitalisierungswahrscheinlichkeit auf der ersten Stichprobe verwendeten Methoden verglichen. Es werden an dieser Stelle nicht alle vorhandenen Modelle, sondern nur die Ergebnisse des jeweils besten Modells für jede Methode miteinander verglichen. In Tabelle 4 sind die Ergebnisse der Modelle zusammenfassend dargestellt.

Tabelle 4: Trefferquoten der besten Modelle auf der ersten Stichprobe

Modellreduktion	Modellreduktion	Anzahl Variablen	Trefferquote (in Prozent)
Entscheidungsbaum	Einzelklassifikation	27	38,79
Zufallswald	einfache durchschnittliche Korrelation	150	40,31
logistische Regression	implementierter Algorithmus	52	40,24

Quelle: eigene Berechnung

Bei jeder Methode wird ein anderer Weg zur Modellreduktion genutzt, jedoch in keinem Fall das volle Modell verwendet. Somit ist eine statistische Variablenreduktion auch nach Anwendung einer Grouper-Software sinnvoll. Die Anzahl der Variablen ist bei dem Zufallswald deutlich höher als bei der logistischen Regression oder dem Entscheidungsbaum. Letzterer weist die mit Abstand geringste Trefferquote von 38,79 Prozent auf und ist sogar schlechter als diejenige der schlechtesten Zufallswälder beziehungsweise logistischen Regression, die in jedem Fall oberhalb von 39 Prozent lag.

Dagegen liegen der Zufallswald und die logistische Regression bezüglich der Trefferquote fast gleichauf. Ein Test auf signifikante Unterschiede fällt trotz der Differenz der Trefferquote des Entscheidungsbaumes zu den Trefferquoten der anderen Methoden negativ aus. Die Ursache hierfür liegt darin begründet, dass der Test die Trefferquoten nicht direkt miteinander vergleicht, sondern nur prüft, inwieweit die beiden Methoden strukturell unterschiedliche Vorhersagen treffen. So werden bei der gleichen Methode und bei verschiedenen Modellen geringere Differenzen der Trefferquoten unter Umständen eher als signifikant verschieden getestet als größere Differenzen bei unterschiedlichen Methoden. Es kann also nicht gesichert gesagt werden, dass eine der drei Methoden in Kombination mit einer Grouper-Software den übrigen beiden überlegen ist, auch wenn die Trefferquoten von Zufallswald und

logistischer Regression auf einen Vorteil gegenüber Entscheidungsbäumen hinweisen.

Zweite Stichprobe: Entscheidungsbäume

Für die zweite Stichprobe mit einem Umfang von lediglich 15.000 Versicherten wird im Gegensatz zur ersten Stichprobe keine Grouper-Software verwendet, sodass darin insgesamt 13.450 Variablen vorhanden sind. Aus diesem Grund ist es zwangsläufig notwendig, die Anzahl der Variablen mittels statistischer Methoden zu reduzieren. Zunächst werden die Ergebnisse der Vorhersage durch Entscheidungsbäume nach dem CART-Algorithmus betrachtet. Die Ergebnisse der sieben erzeugten Entscheidungsbäume zeigt Tabelle 5.

Tabelle 5: Trefferquoten der verschiedenen Entscheidungsbaummodelle auf der zweiten Stichprobe

Modellreduktion	Anzahl Variablen	Trefferquote (in Prozent)
implementierter Algorithmus	380	27,16
Einzelklassifikation	17	30,63
zweifache durchschnittliche Korrelation	1.295	30,41
dreifache durchschnittliche Korrelation	434	32,02
vierfache durchschnittliche Korrelation	182	33,91
fünffache durchschnittliche Korrelation	94	32,75
zehnfache durchschnittliche Korrelation	15	32,58

Quelle: eigene Berechnung

Die Entscheidungsbäume, die nach einer Auswahl der erklärenden Variablen gemäß der Korrelation mit der Zielvariablen erstellt wurden, weisen mit einer Ausnahme höhere Trefferquoten auf als diejenigen Entscheidungsbäume, bei denen die Variablenanzahl mithilfe des implementierten Algorithmus oder mittels der Auswahl nach Einzelklassifikation reduziert wurde. Dabei scheint die Zahl der genutzten Variablen

nicht entscheidend für die Trefferquote zu sein, auch wenn es nicht mehr als wenige Hundert sein sollten. Vielmehr ist auch die Auswahl der spezifischen Variablen, sprich die jeweilige Selektionsmethode, von Bedeutung. Die Ergebnisse von statistischen Tests deuten darauf hin, dass zumindest die Modelle mit 182, 94 und 15 Variablen dabei gleich gut sind, da sie untereinander keine signifikanten Unterschiede aufweisen, während die Unterschiede zu den weiteren Modellen als signifikant gekennzeichnet werden.

Zweite Stichprobe: Zufallswälder

Für die Zufallswälder auf der zweiten Stichprobe werden acht verschiedene Modelle mittels verschiedener Methoden zur Variablenreduktion erzeugt. Die Zahl der Bäume wird hierbei konstant bei 500 belassen. Dabei muss der implementierte Algorithmus zur Auswahl der geeigneten Variablen mehrfach angewendet werden, um eine ausreichend geringe Anzahl von Variablen auszuwählen, sodass diese verarbeitbar sind. In Tabelle 6 sind die Ergebnisse der Zufallswälder dargestellt.

Tabelle 6: Trefferquoten der verschiedenen Zufallswälder auf der zweiten Stichprobe

Modellreduktion	Anzahl Variablen	Trefferquote (in Prozent)
vierfach implementierter Algorithmus	2.381	38,31
zwölfach implementierter Algorithmus	664	37,93
1,6-fache durchschnittliche Korrelation	2.698	37,42
zweifache durchschnittliche Korrelation	1.295	38,91
dreifache durchschnittliche Korrelation	434	39,04
vierfache durchschnittliche Korrelation	182	39,54
fünffache durchschnittliche Korrelation	94	40,04
siebenfache durchschnittliche Korrelation	41	37,94

Quelle: eigene Berechnung

Die zu beobachtenden Differenzen der Trefferquoten sind mit zum Teil mehr als zwei Prozentpunkten erheblich, jedoch deuten Tests in nur wenigen Fällen auf Signifikanz der Unterschiede hin. Es werden lediglich die Zufallswälder mit Trefferquoten unterhalb von 38 Prozent als schlechter als die übrigen herausgestellt. Dennoch deuten die Ergebnisse darauf hin, dass die Zufallswälder mit einer geringeren Zahl von Variablen besser abschneiden, als diejenigen mit einer sehr hohen Variablenzahl. Da aber eine Mindestanzahl von Variablen im Modell vorhanden sein soll, weist das Modell mit 94 Variablen die beste Trefferquote in Höhe von 40,04 Prozent auf, während eine Reduzierung auf 41 Variablen die Trefferquote auf 37,94 Prozent reduziert.

Zweite Stichprobe: logistische Regression

Analog zur ersten Stichprobe werden bei der logistischen Regression lediglich zwei Modelle erzeugt. Die Trefferquote des Modells mit 188 Variablen ist mit 33,27 Prozent deutlich besser als die Trefferquote des Modells mit 1.295 Variablen mit 29,45 Prozent. Dies wird auch durch einen klar positiv ausfallenden Signifikanztest verdeutlicht.

Tabelle 7: Trefferquoten der logistischen Regression auf der zweiten Stichprobe

Modellreduktion	Anzahl Variablen	Trefferquote (in Prozent)
zweifache durchschnittliche Korrelation	1.295	29,45
implementierter Algorithmus	188	33,27

Quelle: eigene Berechnung

Zweite Stichprobe: Vergleich der drei Methoden

Für den Vergleich der drei Methoden zur Vorhersage der Hospitalisierungswahrscheinlichkeiten wird erneut jeweils ein Vertreter jeder Methode ausgewählt. Wegen der großen Anzahl der erzeugten Modelle und um Ausreißer – wie beispielsweise ein zufällig sehr gutes Modell –

auszuschließen, wird für Entscheidungsbäume und Zufallswälder jeweils nur das zweitbeste Modell ausgewählt. Den Vergleich der Trefferquoten der besten Modelle auf der zweiten Stichprobe zeigt Tabelle 8.

Tabelle 8: Trefferquoten der besten Modelle auf der zweiten Stichprobe

Modellreduktion	Modellreduktion	Anzahl Variablen	Trefferquote (in Prozent)
Entscheidungsbaum	vierfache durchschnittliche Korrelation	94	32,75
Zufallswald	fünffache durchschnittliche Korrelation	182	39,54
logistische Regression	implementierter Algorithmus	188	33,27

Quelle: eigene Berechnung

Die Ergebnisse in Tabelle 8 zeigen, dass der Zufallswald mit 39,54 Prozent eine mehr als sechs Prozentpunkte höhere Trefferquote besitzt als die logistische Regression und der Entscheidungsbaum. Bei Letzteren bestehen dagegen keine größeren Unterschiede in der Trefferquote. Die Unterschiede zum Zufallswald sind dabei signifikant, die zwischen Entscheidungsbaum und logistischer Regression hingegen nicht. Soll also eine Grouper-Software nicht angewendet werden, so sind Zufallswälder den Entscheidungsbäumen und der logistischen Regression klar vorzuziehen.

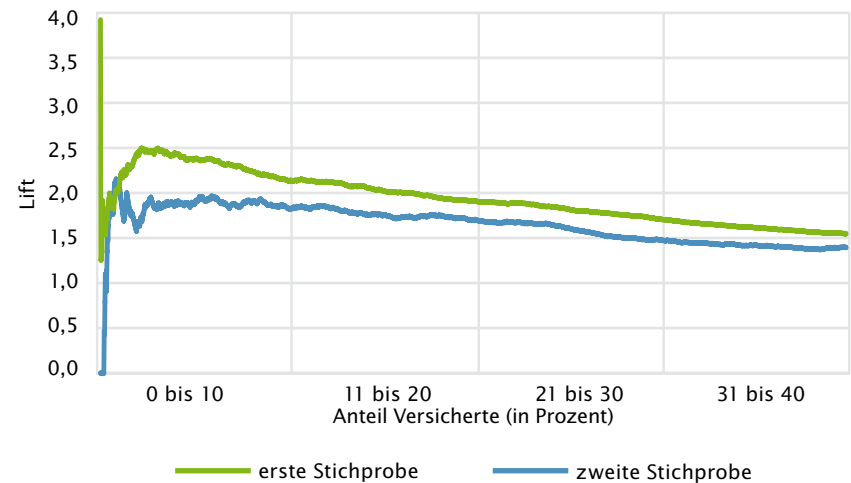
Vergleich der Ergebnisse auf beiden Stichproben

Werden die verschiedenen Methoden zur Wahrscheinlichkeitsberechnung auf den beiden Stichproben verglichen, so kann nicht mehr auf einen Signifikanztest zurückgegriffen werden, da dieser nur bei identischen Testdatensätzen angewendet werden kann. Aus diesem Grund muss auf sogenannte Lift-Grafiken zurückgegriffen werden. Die Lift-Grafiken geben an, um welchen Faktor sich die Trefferquote der jeweiligen Methode von der Trefferquote einer Zufallsauswahl unterscheidet. Dabei werden die Versicherten auf der Abszisse (x-Achse) geordnet nach

der Hospitalisierungswahrscheinlichkeit angegeben und auf der Ordinate (y-Achse) der Lift der Versicherten bis zu diesem Punkt der Abszisse.

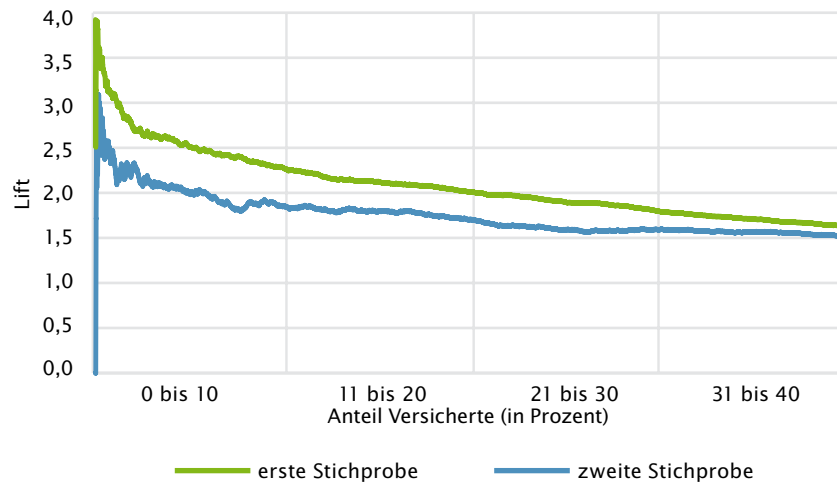
Ein Lift von 2 bei 25 Prozent bedeutet dabei, dass bei den 25 Prozent Versicherten mit der höchsten Hospitalisierungswahrscheinlichkeit doppelt so viele korrekt vorhergesagt werden, wie bei einer zufälligen Vorhersage getroffen würden. Dies bedeutet, dass ein Prognosemodell umso besser ist, je höher der Verlauf des Lift-Charts ist. Die Ursache für den unterschiedlichen Verlauf der Grafiken im niedrigen Prozentbereich ist, dass hier nur wenige Personen in den Wert der Lift-Grafik in dem jeweiligen Punkt eingehen und deswegen zu Beginn vereinzelt falsch klassifizierte Versicherte den Lift stark verringern. Ab einem prozentualen Anteil von ungefähr 0,5 Prozent der Versicherten tritt diese Problematik aufgrund der nun größeren Anzahl von berücksichtigten Personen nicht mehr auf.

Abbildung 1: Lift-Grafik der Entscheidungsbäume mit der besten Trefferquote auf den beiden Stichproben



Quelle: BARMER-Daten 2005/2006

Abbildung 2: Lift-Grafik der logistischen Regressionen mit der besten Trefferquote auf den beiden Stichproben



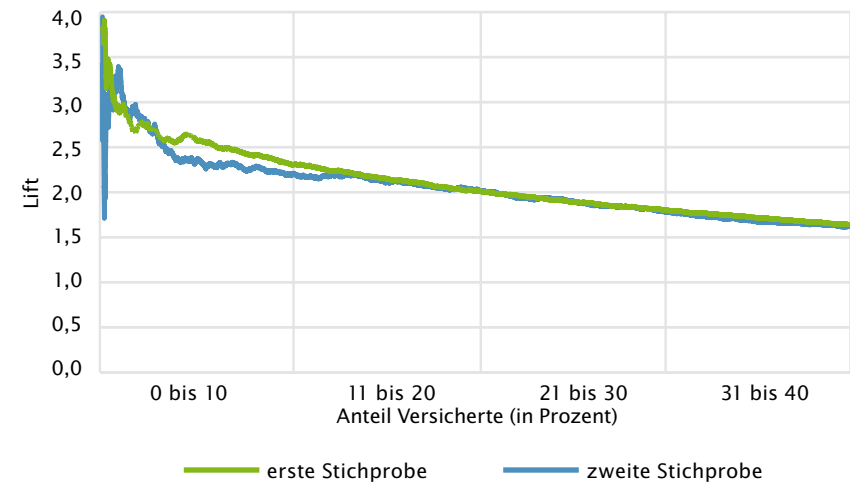
Quelle: BARMER-Daten 2005/2006

Der Vergleich der Methoden auf der Basis der Lift-Grafiken zeigt unterschiedliche Ergebnisse. Sowohl Entscheidungsbäume als auch die logistische Regression liefern auf der ersten Stichprobe deutlich bessere Ergebnisse ab, wie in Abbildung 1 beziehungsweise 2 zu erkennen ist. Sowohl die Entscheidungsbäume als auch die logistische Regression profitieren deutlich von einer Aggregation mithilfe der Grouper-Software.

Bei Entscheidungsbäumen liegt die beste Trefferquote auf der ersten Stichprobe – also mit Nutzung der Grouper-Software – bei 38,79 Prozent, auf der zweiten dagegen bei nur 33,91 Prozent. Noch größer ist sogar der Unterschied im Falle der logistischen Regression. Dort betragen die Trefferquoten 40,24 beziehungsweise 33,27 Prozent. Diese beiden Methoden profitieren somit klar von der Anwendung einer Grouper-Software.

Die Zufallswälder hingegen zeigen, dass ihre Prognosegüte durch die Nutzung einer Grouper-Software nicht mehr weiter gesteigert werden kann und die Trefferquote bereits ohne Grouping auf dem hohen Niveau von 40,04 Prozent liegt. Abbildung 3 veranschaulicht die geringen Prognosegüteunterschiede bei Zufallswäldern mit und ohne Nutzung der Grouper-Software. Der Lift des Zufallswaldes auf der ersten Stichprobe ist nur geringfügig besser als auf der zweiten Stichprobe ohne Nutzung eines Groupers. So unterscheiden sich die Trefferquoten mit 40,31 Prozent auf der ersten und 40,04 Prozent auf der zweiten Stichprobe nur marginal.

Abbildung 3: Lift-Grafik der Zufallswälder mit der besten Trefferquote auf den beiden Stichproben



Quelle: BARMER-Daten 2005/2006

Fazit

Sowohl bei den Entscheidungsbäumen als auch bei der logistischen Regression erlaubt die Anwendung einer Grouper-Software eine deutliche Verbesserung der Vorhersagen. Die Anwendung von einfachen statistischen Methoden zur Reduktion der Anzahl an vorhandenen Variablen stellt hier keine Alternative zum Grouping dar. Soll also eine dieser beiden klassischen Methoden zur Berechnung der Hospitalisierungswahrscheinlichkeit verwendet werden, so ist der Verzicht auf eine Grouper-Software mit einem deutlichen Verlust an Prognosequalität verbunden.

Im Gegensatz zu den genutzten klassischen Verfahren zeigt sich, dass bei der Methode der Zufallswälder eine Grouper-Software zur Voraggregation der Daten nicht notwendig ist. Damit bestätigen die durchgeführten Analysen die in die neue Methode gesetzten Erwartungen. Einzig die gegebenen technischen Restriktionen der genutzten Software verhindern, dass die Zufallswälder ihr volles Potenzial realisieren können und eine signifikante Verbesserung der Prognosegüte erlauben. Daher ist besonders hervorzuheben, dass die hier gezeigte Prognosegüte der Zufallswälder auf einer deutlich kleineren Versichertenanzahl durchgeführt werden musste, als es bei den konkurrierenden Verfahren der Fall war.

Es zeigt sich aber auch, dass das Verfahren der Zufallswälder ebenfalls für die Nutzung bei kleinen Versichertenzahlen geeignet ist und stabile Ergebnisse liefern kann. So haben die klassischen Verfahren große Probleme, bei kleinen Fallzahlen stabile Prognosemodelle zu erzeugen. Dies ist insbesondere bei Versorgungsprogrammen für seltene Krankheiten ein Problem, da hier nur Modellierungen für kleine Versichertenmengen durchgeführt werden können. In diesen Fällen können Zufallswälder ebenfalls sinnvoll angewendet werden.

Somit steht bereits jetzt fest, dass mit der Methode der Zufallswälder eine praktikable Methode zur Prognosemodellierung im Kontext eines Versorgungsmanagements gegeben ist und Kosten, die durch die

Nutzung einer Grouper-Software entstehen, vermeidbar sind. Sofern auf die vorhandenen Vorteile einer Grouper-Software, wie beispielsweise der leichten medizinischen Interpretierbarkeit der Variablen, verzichtet werden kann, steht mit der Methode der Zufallswälder ein Verfahren zur Verfügung, das einen Verzicht auf solche Software ohne Verlust an Prognosequalität ermöglicht.

Perspektivisch ist zu erwarten, dass der Zufallswälder-Algorithmus in leistungsfähige Datenanalyse-Software-Produkte implementiert werden wird. Dies eröffnet die Möglichkeit, mehr Versicherte (also größere Stichproben) betrachten zu können und damit die Prognosegüte weiter zu steigern.

Literatur

- Berk, R. A. (2008): *Statistical learning from a regression perspective*. New York.
- Breiman, L., J. H. Friedman, R. A. Olshen und C. J. Stone (1984): *Classification and Regression Trees*. Boca Raton.
- Breiman, L. (2001): *Random Forests*. *Machine Learning*. Nummer 45 (1). S. 5-32.
- Dietterich, T. G. (1997): *Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms*. Oregon.
- Guyon, I., und A. Elisseeff (2003): *An Introduction to Variable and Feature Selection*. *Journal of Machine Learning Research*. Nummer 3. S. 1157-1183.
- Hosmer, D. W., und S. Lemeshow (1989): *Applied logistic Regression*. New York.
- Shi, T., D. Seligson, A. S. Belldegrun, A. Palotie und S. Horváth (2005): *Tumor Classification by Tissue Microarray Profiling: Random Forest Clustering Applied to Renal Cell Carcinoma*. *Mod Pathol*. Nummer 18 (4). S. 547-57.

- Wolik, A. (2008): Predictive Modeling. Bestimmung von Hospitalisierungswahrscheinlichkeiten mithilfe von Data-Mining-Verfahren. In: U. Repschläger: BARMER Gesundheitswesen aktuell 2008. Beiträge und Analysen zur Gesundheitsreform ab 2009. Wuppertal. S. 286-315.
- Wu, X., Z. Wu und K. Li (2008): Identification of differential gene expression for microarray data using recursive random forest. Chin Med. Nummer 121 (24). S. 2492-2496.
- Yang, F., H. Wang, M. Hong, C. Lin und W. Cai (2009): Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. BMC Bioinformatics. www.ncbi.nlm.nih.gov/pmc/articles/PMC2648734/ (Download 27. Mai 2010).