Methoden zur Prädiktion von Hochnutzern: ein systematischer Literatur-Review

Methods to predict high users: a systematic literature review

Justyna Hartmann[1]
Svenja Schauer[1]
Christian Krauth[1]
Volker Amelung[1]

1 Medizinische Hochschule Hannover, Institut für Epidemiologie, Sozialmedizin und Gesundheitssystemforschung, Hannover, Deutschland

Abstract

Background: A small group of patients accounts for a high amount of health care expenditures in Germany as well as in other countries. A portion of these expenses could be prevented by early identification of potential high users. This is possible through predictive modelling which offers various of methodical approaches. Therefore, the aim of this study is to identify different methodological approaches of predictive modelling of potential high users and to aid the decision-making process for the selection of appropriate method.
Method: A systematic literature search was done in the scientific database SciVerse Scopus in October 2011 and supplemented by a manual search. Two persons selected identified citations in a two-step procedure independently, according to predetermined inclusion and exclusion criteria.
Results: From the 216 identified publications, 18 articles remained after the final selection process. Two different approaches for dealing with this topic can be identified. On the one hand, there is an approach that focuses on patient-characteristics. Therefore, studies using this approach define high cost patients based on the frequency of health care utilization. The methods used for this approach are logistic, linear and negative binomial regression, with logistic regression as the most common one. On the other hand, there is a cost-oriented approach. Papers with this focus are primarily interested in testing different methods and new ways of prediction. The common method of logistic regression is used as well as the very special method of extreme regression. Data-mining techniques and classification systems like diagnostic cost groups are utilized as well. These methods are suitable for preparation and information processing of a large amount of diagnostic data.
Conclusion: Different methods to predict high users exist. The choice of the method depends on the research question, the aim, the data and the available resources. When research focuses on predictors of high usage, logistic regression is a suitable and commonly used method.

Keywords: frequent attenders, high-cost cases, heavy user, predictive modelling, methods

Zusammenfassung

Hintergrund: Auf einen kleinen Anteil von Patienten entfällt ein großer Anteil der Krankheitsausgaben. Dies zeigen sowohl deutsche als auch internationale Studien. Ein Teil dieser Ausgaben könnte durch frühzeitige Identifikation potentieller Hochnutzer vermieden werden. Dies ist unter anderem durch die Entwicklung eines Prädiktionsmodells möglich, wobei die methodische Umsetzung eines solchen Modells sehr unterschiedlich aussehen kann. Ziel dieser Arbeit ist deshalb herauszuarbeiten, welche methodischen Möglichkeiten es gibt, um ein Prädiktionsmodell zu erstellen, mit dem man frühzeitig steuerbare Hochnutzer finden kann, und daraus eine Entscheidungshilfe für die Wahl einer Methode abzuleiten.

Methode: Es wird eine systematische Suchrecherche in der Literaturdatenbank SciVerse Scopus im Oktober 2011 durchgeführt und durch Handrecherche ergänzt. Die Literatur wird in einem zweistufigen Vieraugenverfahren nach vorher definierten Ein- und Ausschlusskriterien selektiert.
Ergebnisse: Von 216 identifizierten Publikationen werden abschließend 18 in die Analysen eingeschlossen. Diese Artikel können in zwei Gruppen unterteilt werden. Auf der einen Seite gibt es Ansätze, bei denen Patienten-Charakteristika im Vordergrund stehen und die Hochnutzer über ihre Inanspruchnahmehäufigkeit von Gesundheitsleistungen definieren. Dabei kommen die Methoden der logistischen, linearen und negativ binomialen Regression zum Einsatz, wobei die logistische die am häufigsten verwendete Methode darstellt. Auf der anderen Seite gibt es kostenorientierte Ansätze. Bei Artikeln mit einem Fokus auf Kosten stehen häufig methodische Aspekte und die Möglichkeiten der Prädiktion im Vordergrund. Die Methode der logistischen Regression kommt ebenso zum Einsatz wie die sehr spezielle Form der „Extreme Regression". Außerdem gibt es mit dem Einsatz von Data-mining Techniken sowie mit Klassifikationssystemen – wie „Diagnostic Cost Groups" – Ansätze, die auf die Aufbereitung und Informationsverarbeitung großer Mengen von Diagnoseinformationen ausgelegt sind.
Schlussfolgerung: Es gibt verschiedene Methoden zur Prädiktion von Hochnutzern. Die Wahl der Methode sollte sich nach der Fragestellung und dem Ziel, der Datenbasis sowie den verfügbaren Ressourcen richten. Bei Ansätzen, bei denen die Wahl geeigneter Prädiktoren im Vordergrund steht, stellt die logistische Regression eine geeignete und häufig verwendete Methode dar.

Schlüsselwörter: Hochnutzer, Hochkostenfälle, Prädiktion, Prädiktionsmodell, Methoden

Hintergrund

Internationale Studien zeigen, dass auf einen kleinen Anteil von Patienten ein großer Anteil der Krankheitsausgaben entfällt. So entfallen in Deutschland im Jahr 2001 etwa 80% der Gesamtausgaben der damaligen Gmünder Ersatzkasse auf nur 10% ihrer Versicherten [1]. Solche Hochnutzer sind Gegenstand zahlreicher nationaler und internationaler Forschungsarbeiten. Vor allem für Kostenträger sind sie relevant, da eine Veränderung der Inanspruchnahme entsprechender Patienten ein hohes Einsparpotential bietet und gleichzeitig die Situation der Patienten verbessern kann. Will man die Eskalation von Krankheitsverläufen vermeiden, ist es notwendig bereits frühzeitig steuernd einzugreifen und den Krankheitsverlauf damit positiv zu beeinflussen. Deutlich wird der Nutzen einer frühen Steuerung am Beispiel der Erkrankung Diabetes mellitus: Wird Typ-2 Diabetes bei Personen mit erhöhtem Risiko frühzeitig erkannt und behandelt, kann durch gezielte Änderung des Lebensstils das Auftreten um über 50% gesenkt werden, während ein spätes oder sehr spätes Reagieren zu lebenslanger Medikamenteneinnahme, Amputationen und weiteren Folgeerkrankungen führen kann [2].
Um gezielt intervenieren zu können und Eskalationen zu vermeiden, ist es deshalb von großem Interesse, zukünftige Hochnutzer frühzeitig zu identifizieren. Dies kann unter anderem durch die Entwicklung eines Prädiktionsmodells erreicht werden. Prädiktoren sind Merkmals-Variablen, mit denen sich ein möglicher kausaler Einfluss auf ein abhängiges Merkmal modellieren und überprüfen lässt [3]. Ein Prädiktionsmodell dient folglich dazu, kausale Beziehungen zu identifizieren und den Einfluss von unabhängigen Variablen auf ein Merkmal zu prognostizieren. Die methodische Umsetzung eines solchen Modells kann dabei sehr unterschiedlich aussehen. Ziel dieser Studie ist deshalb, mithilfe einer systematischen Literaturrecherche herauszuarbeiten, welche methodischen Möglichkeiten es gibt, um ein Prädiktionsmodell zu erstellen mit dem man frühzeitig steuerbare Hochnutzer finden kann, und Entscheidungshilfen für die Wahl einer Methode abzuleiten.

Methode

Eine systematische Suchrecherche wird am 12.10.2011 in der Literaturdatenbank SciVerse Scopus, einer Abstract- und Zitationsdatenbank mit aktuell 18.500 internationalen Titeln, durchgeführt und durch eine Handrecherche ergänzt. Die Suchrecherche besteht aus einer Verknüpfung von Begriffen für Hochnutzer mit Begriffen, die für Prädiktionsmodelle verwendet werden (siehe Abbildung 1). Bei den Begriffen für Hochnutzer wird keine Einschränkung hinsichtlich der dahinterstehenden Definition festgelegt, so dass eine Bandbreite von unterschied-

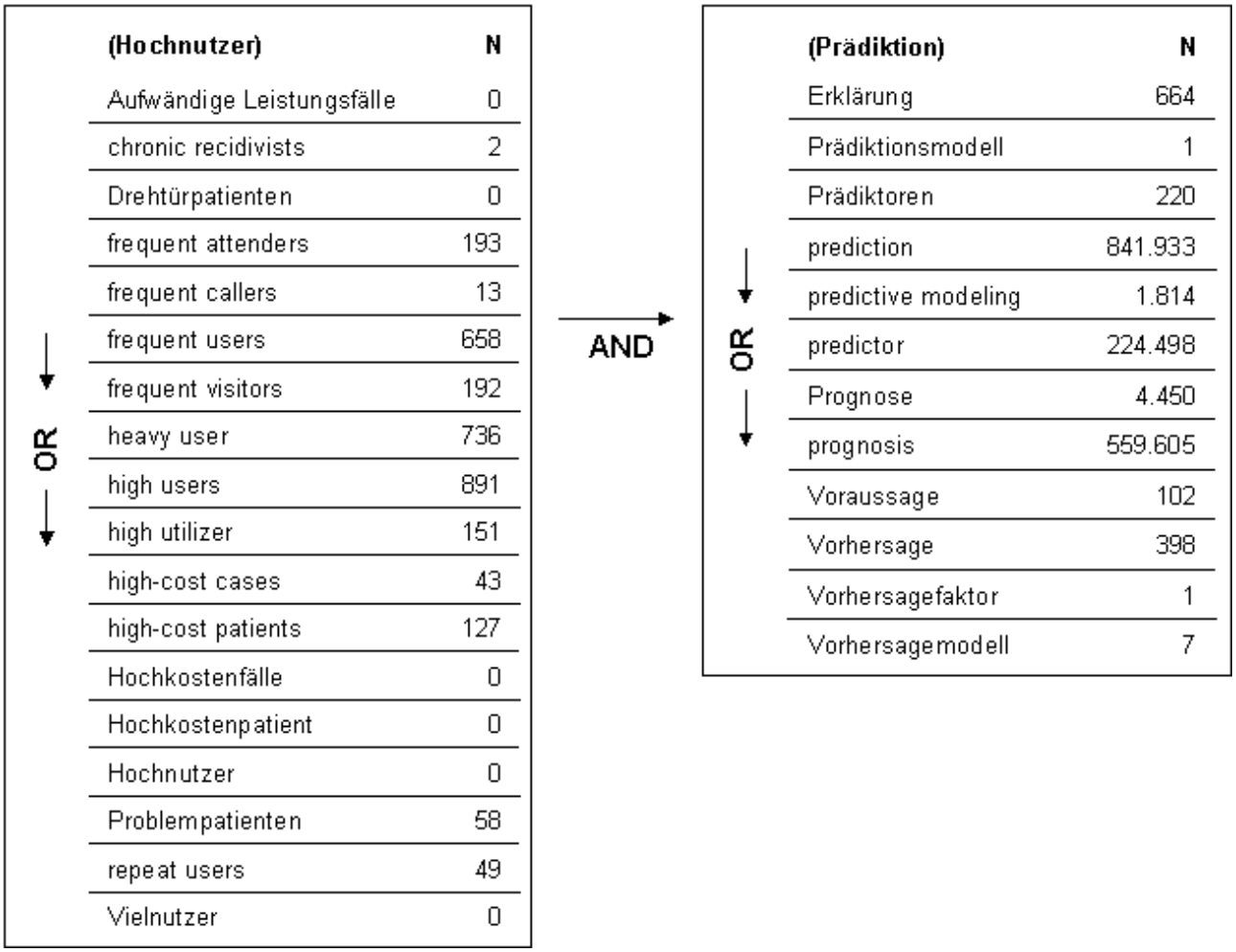

Abbildung 1: Suchstrategie

lich definierten Hochnutzer-Begriffen eingeschlossen wird.

Sprachliche Einschränkungen werden nicht vorgenommen. Die so erhaltene Literatur wird in einem zweistufigen Vieraugenverfahren von zwei unabhängigen Gutachtern gesichtet. In einem ersten Schritt wird die Literatur anhand des Titels und des Abstracts gemäß zuvor definierter Ein- und Ausschlusskriterien selektiert:

Einschlusskriterien:

- Zeit: Literatur der letzten 10 Jahre
- Endpunkt: Hochnutzer
- Art der Publikation: methodische Artikel, Studien mit quantitativen Ergebnissen
- Qualität: Beschreibung der verwendeten Methodik

Ausschlusskriterien:

- Endpunkt: Artikel, bei denen nicht primär Hochnutzer im Vordergrund stehen
- Artikel, die sich ausschließlich auf dauerhafte Hochnutzer fokussieren
- Art der Publikation: Meinungen, Kommentare, qualitative Studien

Anschließend werden die verbleibenden Studien im Volltext bestellt und in einem zweiten Schritt entsprechend der Ein- und Ausschlusskriterien hinsichtlich ihrer Relevanz für das vorliegende Thema geprüft und ggf. ausgeschlossen. Die verbleibenden Artikel werden analysiert und die Ergebnisse in Kategorien zusammengefasst beschrieben. Dazu gehören die Länder, in denen die Forschung stattgefunden hat, die Studienpopulationen, die Untersuchungsgegenstände, die Datenbasen, die verwendeten Methoden und Modelle und die Vor- und Nachteile dieser Methoden. Abschließend werden die Ergebnisse in Form von Bewertungsparametern in einer Tabelle zusammengefasst.

Ergebnisse

Insgesamt werden 216 Artikel mithilfe der oben genannten Suchstrategie identifiziert (siehe Abbildung 2). Nach Durchsicht von Titeln und Abstracts werden 193 Artikel ausgeschlossen. Die verbleibenden 23 Artikel sowie ein durch Handrecherche identifizierter Artikel finden im zweiten Selektionsschritt Berücksichtigung. Nach Durchsicht der Volltexte verbleiben 18 Artikel [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21]. Diese bilden die Grundlage für die folgenden Analysen.

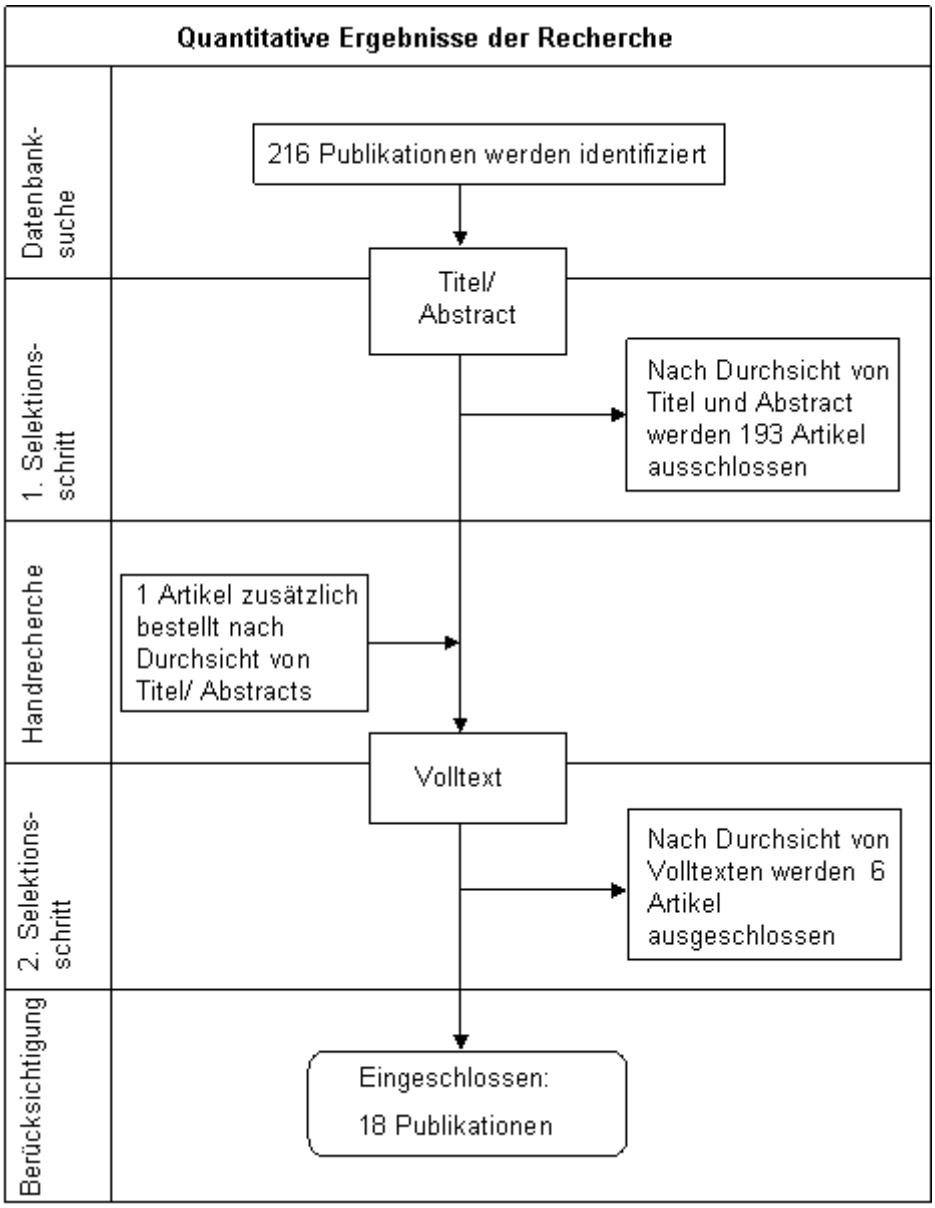

Abbildung 2: Flussdiagramm der Studienselektion

Länder

Es zeigt sich, dass in den Vereinigten Staaten von Amerika (USA) vor allem eine kostenorientierte Perspektive dominant ist (siehe Tabelle 1). Acht der 18 Artikel stammen aus den USA und sieben davon legen ihren Fokus auf methodische Aspekte bzw. auf Modellierungen, die auf Kosten ausgelegt sind und weniger auf Charakteristika von Hochnutzern [5], [8], [14], [15], [16], [20], [21]. Die USA ist die am häufigsten vertretene Nation in der Recherche. Deutschland und Groß-Britannien folgen mit je zwei Artikeln. Alle übrigen Nationen sind mit je einem Artikel vertreten, wobei Artikel aus Südamerika über Vorderasien bis Europa vorhanden sind. Das deutet darauf hin, dass das Thema der Hochnutzung ein weit verbreitetes Phänomen ist, das nicht etwa an ein bestimmtes Gesundheitssystem gebunden ist.

Studienpopulation

Insgesamt beziehen sieben der 18 Artikel ihre Untersuchungen auf eine bestimmte Subpopulation (siehe Tabelle 1). Neben einem Fokus auf ältere Menschen [16] werden vor allem bestimmte Erkrankungen ausgewählt, die bei den Hochnutzern vorhanden sein müssen [6], [7], [9], [17], [18], [20]. Es handelt sich dabei meist um Erkrankungen, die häufig auftreten, aber sehr unterschiedliche Verläufe aufweisen können, was zu hohen Kostendifferenzen führt, wie z.B. die chronisch obstruktive Lungenerkrankung (COPD) [20], Diabetes [6], [9] oder Schizophrenie [17].

Untersuchungsgegenstand

Der Ausgangspunkt der Artikel ist bei allen derselbe: Eine geringe Anzahl von Patienten lässt sich im Vergleich mit der restlichen Untersuchungspopulation als Hochnutzer

Tabelle 1: Länder, Studienpopulationen und Untersuchungsgegenstände aller einbezogenen Studien

Nr.	Autor	Land	Studienpopulation	Untersuchungsgegenstand
4	Al-Kandari et al. 2008	Kuwait	Gesamtpopulation	charakteristikorientiert
5	Ash et al. 2001	USA	Gesamtpopulation	kostenorientiert
6	Botica et al. 2007	Kroatien	Subgruppe: Erkrankung	charakteristikorientiert
7	Etemad, McCollam 2005	USA	Subgruppe: Erkrankung	charakteristikorientiert
8	Fleishman, Cohen 2010	USA	Gesamtpopulation	kostenorientiert
9	Gregori et al. 2009	Italien	Subgruppe: Erkrankung	kostenorientiert
10	Kapur et al. 2004	UK	Gesamtpopulation	charakteristikorientiert
11	Kersnik, Svab, Vegnuti 2001	Slowenien	Gesamtpopulation	charakteristikorientiert
12	Kobus et al. 2009	Brasilien	Gesamtpopulation	kostenorientiert
13	Little et al. 2001	UK	Gesamtpopulation	charakteristikorientiert
14	Moturu, Liu, Johnson 2008a	USA	Gesamtpopulation	kostenorientiert
15	Moturu, Liu, Johnson 2008b	USA	Gesamtpopulation	kostenorientiert
16	Rakovski et al. 2002	USA	Subgruppe: Ältere Menschen	kostenorientiert
17	Roick et al. 2004	Deutschland	Subgruppe: Erkrankung	charakteristikorientiert
18	Spießl et al. 2002	Deutschland	Subgruppe: Erkrankung	charakteristikorientiert
19	Vedsted et al. 2001	Dänemark	Gesamtpopulation	charakteristikorientiert
20	Yarger et al. 2008	USA	Subgruppe: Erkrankung	kostenorientiert
21	Zhao et al. 2003	USA	Gesamtpopulation	kostenorientiert

bezeichnen. Zudem liegt allen Artikeln die Idee zugrunde, dass Hochnutzer, die früh genug identifiziert werden, steuerbar sind und so hohe Kosten vermieden werden können. Dabei kann der Fokus darauf liegen, dass eine geringe Anzahl von Patienten unverhältnismäßig hohe Ausgaben, einen unproportional hohen Arbeitsaufwand seitens der Ärzte oder eine überdurchschnittliche Inanspruchnahme von Gesundheitsleistungen verursacht. Das Ziel besteht darin, ein Prädiktionsmodell zu entwickeln bzw. verschiedene Modelle zu testen und Methoden zu finden, um diese Hochnutzer möglichst früh und genau vorhersagen zu können.

Die Studien lassen sich entsprechend ihres Studiengegenstandes in zwei Gruppen einteilen (siehe Tabelle 1): Die einen orientieren sich an Kosten und die anderen an Patientencharakteristika. In der ersten Gruppe besteht das Interesse der genauen Vorhersage primär darin, die Kosten im Gesundheitssystem möglichst präzise zu prognostizieren [5], [8], [9], [12], [14], [15], [16], [20], [21]. Was die genauen Ursachen für die hohen Ausgaben sind und wodurch sich Hochnutzer auszeichnen, spielt keine

nennenswerte Rolle, selbst wenn neben der Vorhersage der Kosten die Identifizierung für die Implementierung eines Case Managements als Ziel genannt wird. In der zweiten Gruppe steht die Suche nach Charakteristika, die Hochnutzer ausmachen und die das Nutzungsverhalten beeinflussen, im Vordergrund [4], [6], [7], [10], [11], [13], [17], [18], [19]. Es geht in erster Linie darum, geeignete Prädiktoren zu finden, die zu einer häufigen Inanspruchnahme von Gesundheitsleitungen führen.
Der primäre Unterschied zwischen diesen beiden Gruppen besteht also darin, dass kostenorientierte Modelle auf die Vorhersage abzielen, während die charakteristikorientierten Modelle stärker das „Warum?“ betonen und den Fokus damit auf die Steuerung legen.

Datenbasis

Die unterschiedlichen Herangehensweisen in den Artikeln stellen unterschiedliche Anforderungen an die Datenbasen (siehe Tabelle 2). Auf Kosten ausgerichtete Studien müssen Zugriff auf Kostendaten haben und Untersuchungen, die Diagnosen gruppieren, müssen über Diagnosedaten verfügen. In den kostenorientierten Artikeln wird deshalb auf große Datenbanken zurückgegriffen. Insgesamt kommen in allen kostenorientierten Artikel Sekundärdaten zum Einsatz [5], [8], [9], [12], [14], [15], [16], [20], [21]. Primärdaten sind vor allem in der Gruppe der an Charakteristika interessierten Artikel in Gebrauch, da Sekundärdaten häufig nicht die für diese Fragestellungen relevanten Variablen (in ausreichender Qualität) enthalten. So werden von den charakteristikorientierten Studien lediglich drei ausschließlich mit Sekundärdaten [6], [7], [18], eine mit Sekundärdaten und zusätzlich erhobener Daten [4] und fünf mit Primärdaten durchgeführt [10], [11], [13], [17], [19]. Die Primärdaten beinhalten unterschiedliche Informationen, die je nach Fragestellung per Interview oder Fragebogen durch die Patienten oder ihre Ärzte, teilweise zu mehreren Zeitpunkten, erhoben wurden.

Methoden und Modelle

Neben der Unterscheidung zwischen kosten- und charakteristikorientierten Studien kann eine weitere Unterteilung zwischen den Artikeln gemacht werden, die ihren Umgang mit Modellen betrifft. Einige fokussieren methodische Aspekte der Modellfindung und andere legen mehr Wert auf die einzelnen Prädiktoren (siehe Tabelle 2). Die Herangehensweise ist dementsprechend vielseitig und wird daher für jede Studie einzeln dargestellt.

Methodisch orientiertes Modell mit logistischer Regression: Vedsted et al.

Vedsted et al. [19] untersuchen, wie gut zwei Indizes psychologischer Selbsteinschätzungs-Fragebögen als Prädiktoren geeignet sind, um Hochnutzer vorherzusagen. Sie führen dazu eine logistische Regression mit der dichotomen Ausprägung „Hochnutzer/Nicht-Hochnutzer“ als abhängige Variable durch.

Methodisch orientiertes Modell mit „Extreme Regression“: Gregori et al.

Gregori et al. [9] setzen sich mit der Problematik auseinander, dass die abhängige Variable bei Hochnutzern heterogen und asymmetrisch ist, was die Anwendung von Methoden, die auf Normalverteilung ausgerichtet sind, problematisch macht. Sie testen deshalb anhand von drei Beispieldatensätzen, ob sich die Methode der „Extreme Regression“ für die Prädiktion von Hochnutzern eignet.

Methodisch orientiertes Modell mit Data-mining Techniken: Kobus et al.

Kobus et al. [12] versuchen mithilfe von Data-mining Techniken zum einen Beziehungen zwischen Ereignissen zu finden, die zu Gesundheitsleistungsinanspruchnahme führen. Zum anderen testen sie mithilfe des Software-Systems Weka verschiedene Methoden, mit denen man Hochnutzer vorhersagen kann. Sie vergleichen dabei die Methoden J48-Algorithmus, Bagging und AdaBoosting.

Methodisch orientiertes Modell mit Data-mining Techniken: Moturu, Liu und Johnson und Moturu, Liu und Johnson

Bei den beiden Artikeln von Moturu, Liu und Johnson [14], [15] geht es ebenfalls um die Anwendung von Data-mining Techniken. Beide Artikel beziehen sich auf dieselbe Untersuchung, bei der sie die Weka-Software zur Vorhersage zukünftiger Hochnutzer unter Einbeziehung von über 400 Variablen nutzen. Als Prädiktionsmethoden werden dabei AdaBoost, LogitBoost, logistische Regression, Logistic Model Trees und Support Vector Machine verglichen.

Methodisch orientiertes Modell mit Diagnostic-Cost-Group-Klassifikation: Ash et al., Zhao et al. und Yarger et al.

Die Autoren Ash et al. [5], Zhao et al. [21] und Yarger et al. [20] beschäftigen sich mit Klassifikationssystemen, mit denen die zahlreichen ICD-Codierungen aufbereitet und für die Vorhersage von Hochnutzern nutzbar gemacht werden können. Sie verwenden dafür die Software DxCG, die eigens dafür entwickelt wurde. Welche Prädiktionsmethode angewendet wird, wird allerdings nicht genannt, da die Anwendbarkeit der Klassifikationsmethoden im Vordergrund steht. Ash et al. [5] testen unter Einbeziehung von Alter und Geschlecht, ob sich „Diagnostic Cost Group Hierarchical Condition Categories“ (DCG/HCC) besser zur Vorhersage der 0,5% teuersten Patienten eignet als die Vorhersage mithilfe der Kosten aus dem Vorjahr. Zhao et al. [21] prüfen, ob sich die 0,5% teuersten Patienten mit den Vorjahrskosten, mit Alter, Ge-

Tabelle 2: Datenbasen, Modelle und Methoden aller einbezogenen Studien

Nr.	Autor	Datenbasis	Modelle	Methoden
4	Al-Kandari et al. 2008	Primär- und Sekundärdaten	Prädiktoren-orientiert	Logistische Regression
5	Ash et al. 2001	Sekundärdaten	methodisch orientiert	DCG-Klassifikation
6	Botica et al. 2007	Sekundärdaten	Prädiktoren-orientiert	Logistische Regression
7	Etemad, McCollam 2005	Sekundärdaten	Prädiktoren-orientiert	Logistische Regression
8	Fleishman, Cohen 2010	Sekundärdaten	methodisch/ Prädiktoren orientiert	Logistische Regression
9	Gregori et al. 2009	Sekundärdaten	methodisch orientiert	Extreme Regression
10	Kapur et al. 2004	Primärdaten	Prädiktoren-orientiert	Negativ binomiale Regression
11	Kersnik, Svab, Vegnuti 2001	Primärdaten	Prädiktoren-orientiert	Logistische Regression
12	Kobus et al. 2009	Sekundärdaten	methodisch orientiert	Data-mining
13	Little et al. 2001	Primärdaten	Prädiktoren-orientiert	Logistische Regression
14	Moturu, Liu, Johnson 2008a	Sekundärdaten	methodisch orientiert	Data-mining
15	Moturu, Liu, Johnson 2008b	Sekundärdaten	methodisch orientiert	Data-mining
16	Rakovski et al. 2002	Sekundärdaten	methodisch orientiert	Logistische Regression + Klassifikation
17	Roick et al. 2004	Primärdaten	Prädiktoren-orientiert	Logistische Regression
18	Spießl et al. 2002	Sekundärdaten	Prädiktoren-orientiert	Lineare Regression
19	Vedsted et al. 2001	Primärdaten	methodisch orientiert	Logistische Regression
20	Yarger et al. 2008	Sekundärdaten	methodisch orientiert	DCG-Klassifikation
21	Zhao et al. 2003	Sekundärdaten	methodisch orientiert	DCG-Klassifikation

schlecht und DCG oder mit einer Kombination aus allem am besten vorhersagen lassen. Yarger et al. [20] testen indes die Vorhersagekraft der Vorjahreskosten sowie von drei DCG Varianten auf die 5% teuersten Patienten.

Methodisch orientiertes Modell mit Klassifikation und logistischer Regression: Rakovski et al.

Rakovski et al. [16] untersuchen die Vorhersageleistung für in Pflege verbrachter Tage. Sie testen mittels logistischer Regression die Prädiktionsfähigkeit der in Pflege verbrachten Tage des Vorjahrs, der „Adjusted Diagnostic Groups“ (ADG/HCC) sowie eines Modell, das beide Komponenten enthält. In allen drei Fällen wird für Alter und Geschlecht adjustiert.

Methodisch und Prädiktoren orientiertes Modell mit logistischer Regression: Fleishman und Cohen

Fleishman und Cohen [8] haben ebenfalls DCG in ihre Untersuchung einbezogen. Sie haben mittels logistischer Regression allerdings zehn verschiedene Modelle mit unterschiedlichen Komponenten getestet, so dass sie die Schnittstelle zu Artikeln bilden, bei denen die Patientencharakteristika stärker im Vordergrund stehen und nicht methodische Aspekte.

Prädiktoren orientiertes Modell mit logistischer Regression: Al-Kandari et al., Kersnik, Svab und Vegnuti und Little et al.

Den Fokus auf Prädiktoren legen Al-Kandari et al. [4], Kersnik, Svab und Vegnuti [11] sowie Little et al. [13]. Sie versuchen jeweils mittels logistischer Regression Faktoren zu identifizieren, die einen positiven Einfluss darauf haben, dass jemand zu einem Hochnutzer wird. Während Al-Kandari et al. [4] und Little et al. [13] sich ausschließlich auf Patienteneigenschaften fokussieren, beziehen Kersnik, Svab und Vegnuti [11] in ihre Untersuchung sowohl Patienten- als auch Arzt-Eigenschaften mit ein.

Prädiktoren orientiertes Modell mit logistischer Regression in einer Subgruppe: Botica et al. und Etemad und McCollam

Botica et al. [6] und Etemad und McCollam [7] konzentrieren sich in ihren Analysen auf bestimmte Krankheits-Subgruppen. Sie wenden dabei logistische Regressionen an. Botica et al. [6] beschäftigen sich mit Diabetes-Patienten und wählen entsprechende für die Krankheit relevante Prädiktoren. Etemad und McCollam [7] wählen die Subgruppe der Patienten mit akutem Koronarsyndrom (ACS) und versuchen Faktoren zu identifizieren, die zu hoher Nutzung teurer Gesundheitsleistungen führen.

Prädiktoren orientiertes Modell in der Subgruppe mentaler Erkrankungen mit logistischer Regression: Roick et al.

Roick et al. [17] beschäftigen sich in ihren Analysen mit der Subgruppe mentaler Erkrankungen und beziehen entsprechend speziell für diesen Bereich wichtige Faktoren in ihre Analyse ein. Sie untersuchen mittels logistischer Regression Prädiktoren für häufige Einweisungen schizophreniekranker Patienten.

Prädiktoren orientiertes Modell in der Subgruppe mentaler Erkrankungen mit linearer Regression: Spießl et al.

Spießl et al. [18] hingegen nutzen in ihrer Analyse im psychiatrischen Bereich lineare Regression mit einer metrischen abhängigen Variable und prüfen den Effekt bestimmter Prädiktoren auf die Anzahl stationärer Aufenthalte bzw. auf die kumulierte stationäre Behandlungsdauer.

Prädiktoren orientiertes Modell mit negativ binomialer Regression: Kapur et al.

Kapur et al. [10] verwenden ebenfalls eine metrische abhängige Variable, indem sie mittels schrittweiser negativ binomialer Regression psychosoziale und krankheitsbezogene Faktoren für die Besuchshäufigkeit einer Arztpraxis auf ihre Vorhersagekraft hin testen. Zusätzlich wird eine longitudinale Analyse durchgeführt um den Effekt über fünf Jahre hinweg zu messen.

Vor- und Nachteile der Methoden

Ausgehend von der Frage, welche Methoden für die Prädiktion zukünftiger Hochnutzer zur Verfügung stehen, lassen sich die Ergebnisse wie folgt zusammenfassen.

„Diagnostic Cost Goups“ und andere Klassifikationssysteme

Streng genommen handelt es sich bei solchen Klassifikationssystemen nicht um eine Prädiktionsmethode. Vielmehr dienen sie der Aufbereitung und Nutzbarmachung großer Mengen an Diagnose- und Kosteninformationen, die wiederum als Prädiktoren für Hochnutzer verwendet werden. Durchgeführt werden diese Klassifikationen mithilfe spezieller Softwaresysteme, die eigens dafür entwickelt wurden. Cucciare und O'Donohue [22] (die einen guten Überblick über die Methode und ihre Vor- und Nachteile geben) weisen darauf hin, dass die Implementierung eines solchen Systems nicht nur eine sehr aufwendige und komplizierte Angelegenheit ist, sondern dass auch die Kosten dafür sehr hoch sind. Betrachtet man die Modellgüte-Werte, die bei Zhao et al. [21] und Yarger et al. [20] berichtet werden, so zeigt sich, dass mit solchen Modellierungen eine Varianzaufklärung von max. 21% erreicht werden kann, was keine Überlegenheit dieser Modellierungsmethode gegenüber anderen Modellen darstellt.

Data-mining Techniken

Data-mining Techniken dienen dazu, Muster und Strukturen in großen Datenmengen zu finden. Dabei kommen verschiedene Methoden zum Einsatz. In den einbezogenen Studien liegt der Fokus auf bestimmten Algorithmen, die eingesetzt werden können (zu genaueren Informationen zu den Algorithmen siehe [12], [14], [15]. Kobus et al. [12] finden dabei, dass die Methode des Boosting die größte Genauigkeit liefert, während Moturu, Liu und Johnson [14], [15] bei ihren fünf geprüften Techniken keine nennenswerten Unterschiede feststellen können. Der Vorteil von Data-mining besteht darin, dass man große Datenmengen einbeziehen kann und man die potentiellen Zusammenhänge nicht im Vorfeld definieren muss. Ein Nachteil ist ihre hohe Komplexität in der Bedienung und die Unübersichtlichkeit über die genutzten Prädiktoren, da auch hier eine große Vielzahl von Variablen verwendet wird. Je größer das Wissen über das zu untersuchende Feld ist, desto mehr stellt sich deshalb die Frage, ob man mit einigen gezielt ausgewählten Variablen nicht ähnliche Modellerfolge erzielen kann wie unter Einbeziehung aller Patientendaten.

„Extreme Regression“

Bei dem „Extreme Regression“ Ansatz (für mehr Informationen siehe [23]) handelt es sich um eine sehr spezielle und wenig verbreitete Methode. Sie bezieht die Idee von Baumdiagrammen mit ein, ohne dabei aber die Glätte der Funktion zu verlieren. Sie ist besonders gut geeignet, um extreme Outcome-Gruppen zu beschreiben und eignet sich auch für kleine Fallzahlen. In ihrer Funktionsweise ähnelt diese Methode der logistischen Regression, ist aber besser geeignet, wenn es darum geht eine kleine Anzahl kontinuierlicher Prädiktoren zu betrachten [23]. Der Vorteil dieser Methode besteht zudem darin, dass sie gut geeignet ist mit schiefen Verteilungen umzugehen und zudem das Problem umgeht, im Vorfeld einen Cut-off Punkt wählen zu müssen, der die Grenze zwischen Hochnutzern und Vergleichsgruppe markiert [9].

Logistische, lineare und negativ binomiale Regression

Mittels Regressionen können Zusammenhänge unabhängiger Prädiktoren auf eine abhängige Variable überprüft werden (für eine Einführung in die Regressionsanalyse siehe [24]). Während die lineare Regression eine kontinuierliche abhängige Variable voraussetzt, erfordert die logistische eine dichotome und die negativ binomiale eine absolute Variable. Die logistische Regression kam in den betrachteten Artikeln am häufigsten zum Einsatz. Ein Vergleich von Rakovski et al. [16] zeigt, dass sie besser geeignet ist als die lineare Regression. Andererseits wird bei Spießl et al. [18] eingewendet, dass die Nutzung einer kontinuierlichen abhängigen Variable und damit auch der linearen Regression angemessener ist, da kein eindeutiger Punkt identifizierbar ist, der eine klare Trennlinie zwischen Hochnutzern und Nicht-Hochnutzern markieren würde. Das Problem der linearen Regression, obwohl sie den Vorteil bietet keinen Cut-off Punkt bestimmen zu müssen, besteht allerdings darin, dass die abhängige Variable theoretisch alle Ausprägungen von $-\infty$ bis $+\infty$ annehmen können muss. Die Modellgüte-Werte der betrachteten Studien variieren stark je nach einbezogenen Variablen und der Untersuchungspopulation. Die höchste Varianzerklärung von 68% findet sich in einem Modell mit logistischer Regression in einer Subgruppe von Diabetes-Patienten [6].

Diskussion und Schlussfolgerung

Aufgrund der Heterogenität der Studien scheint eine abschließende Bewertung wenig zielführend. Zum einen werden nicht in allen Studien die Modellgüte-Werte berichtet und zum anderen unterscheiden sich die Modellgüte-Maße je nach Methode. Außerdem unterscheiden sich auch die Modellgüte-Werte derselben Methode, da der Erfolg eines Modells maßgeblich von den einbezogenen Variablen abhängt. Die Wahl der Methode selbst richtet sich deshalb nach den einbezogenen Variablen und dem verwendetem Datensatz. Das Skalenniveau der Variablen sowie die Verfügbarkeit und Qualität der Daten tragen zur Wahl der Methode ebenso bei wie die verfügbaren finanziellen und zeitlichen Ressourcen. Außerdem spielt die Art der Fragestellung bei der Wahl der angemessenen Methode eine wichtige Rolle. In Tabelle 3 werden die entscheidungsrelevanten Punkte für jede Methode zusammenfassend dargestellt, so dass Tabelle 3 als Entscheidungshilfe bei der Wahl einer Methode genutzt werden kann.

Bei den kostenorientierten Studien kommen sehr unterschiedliche methodische Ansätze und verschiedene Herangehensweisen zum Einsatz. Deshalb ist bei dieser Art von Studien immer individuell zu entscheiden, welche Methode verwendet werden soll. Im Bereich der charakteristikorientierten Studien ist die Entscheidung für eine Methode zwar auch immer individuell zu treffen, doch zeigt sich in diesem Bereich eindeutig, dass die logistische Regression die am häufigsten verwendete Methode zur Prädiktion von Hochnutzern darstellt. Auch wenn es erforderlich ist die abhängige Variable zu dichotomisieren und damit bewusst eine Grenze zu schaffen, ab der aus einem „Normalnutzer“ ein Hochnutzer wird, stellt die logistische Regression eine geeignete Methode dar, da sie wenig voraussetzungsvoll ist im Vergleich zu anderen Methoden und vielen Datenstrukturen gut genügen kann. Es gibt zwar Methoden die der schiefen Verteilung und dem häufigen Vorkommen von Nullwerten besser gewachsen sind (wie z.B. „Extreme Regression“), aber diese sind auch ungleich komplexer und aufwendiger. Im Vergleich zu den übrigen Möglichkeiten bietet die logistische Regression den methodisch einfachsten Zugang, da weder spezielle Softwaresysteme notwendig sind, noch aufwendig erstellte Klassifikationen eine Rolle spielen, die die Interpretation einzelner Prädiktoren erschweren.

Die Ergebnisse haben zudem gezeigt, dass es lohnenswert sein kann bei der Entwicklung eines Prädiktionsmodells zur frühzeitigen Identifikation von Hochnutzern auch in Betracht zu ziehen, nach Krankheiten differenzierte Subgruppen zu bilden, da daraufhin eine präzisere und treffgenauere Auswahl an einflussreichen Prädiktoren getroffen werden kann.

Limitationen

Die Ergebnisse der systematischen Suchrecherche könnten durch die Wahl der genutzten Datenbank, die möglicherweise nicht alle relevanten Artikel enthält, limitiert sein. Allerdings beinhaltet die gewählte Datenbank SciVerse Scopus sämtliche Titel aus den häufig verwendeten Medline-, Springer- und Elsevier-Datenbanken, sowie aus zahlreichen weiteren. Außerdem ist es möglich, dass interessante Beiträge über Prädiktionsmethoden durch die Stichwortsuche nicht entdeckt werden, da ihr Schwerpunkt nicht methodischer Natur ist, wenngleich die Handrecherche diesem Problem entgegenwirkt. Da nicht in allen Studien die Methoden im Vordergrund stehen, werden auch nicht in allen Artikeln Angaben zur Modellgüte gemacht, was es erschwert, die Ergebnisse zu vergleichen. Außerdem wurden die verschiedenen

Tabelle 3: Bewertungsparameter für die Methoden

Verfahren	Verwendet für	Vorraussetzung	Anforderungen an Daten	Software	Zeitlicher Aufwand
Logistische Regression	Kosten und Charakteristika	Prädiktoren müssen bekannt sein	Abhängige Variable muss dichotom skaliert sein	handelsübliche Analysesoftware (Bsp.: SPSS, Stata, R)	gering
Lineare Regression	Charakteristika	Prädiktoren müssen bekannt sein	Abhängige Variable muss metrisch skaliert sein	handelsübliche Analysesoftware (Bsp.: SPSS, Stata, R)	gering
Negativ binomiale Regression	Charakteristika	Prädiktoren müssen bekannt sein	Abhängige Variable muss absolut skaliert sein	handelsübliche Analysesoftware (Bsp.: SPSS, Stata, R)	gering
Extreme Regression	Kosten	Prädiktoren müssen bekannt sein	keine Information verfügbar	keine Information verfügbar	keine Information verfügbar
Klassifikations system	Kosten	es müssen keine einzelnen Prädiktoren benannt werden	Diagnose-/Kostendaten notwendig	spezielle Klassifikations-Software	hoch
Data-mining Techniken	Kosten	explorativ	große Datenmengen notwendig	spezielle Data-mining Software (Bsp.: Weka)	hoch

Studien an sehr unterschiedlichen Datensätzen überprüft, die – trotz teilweise gleicher Variablenauswahl – durchaus auch unterschiedliche Ergebnisse hinsichtlich ihrer Modellgüte hervorgebracht haben [25].

Ausblick

In Deutschland ist es vor allem für Krankenkassen interessant, Interventionen für Hochnutzer zu entwickeln, die die Versorgung der Patienten verbessern und die Ausgaben der GKV senken können. Um möglichst gezielte Interventionen durchführen zu können, ist es wichtig die Faktoren zu kennen, die die Hochnutzung bewirken. Als Datenquelle stehen Sekundärdaten in Form von GKV-Routinedaten zur Verfügung, die zeitnah und kostengünstig für solche Zwecke nutzbar gemacht werden könnten. Studien aus anderen Ländern haben gezeigt, dass charakteristikorientierte Ansätze zumeist mit Primärdaten durchgeführt werden, mit dem Argument, dass Sekundärdaten nicht alle relevanten Prädiktoren enthalten. Aus diesem Grund wäre es interessant zu prüfen, ob dies auch für die in Deutschland zur Verfügung stehenden GKV-Routinedaten gilt, oder ob diese für die Entwicklung eines Prädiktionsmodells geeignet sind und ob sich die vergleichsweise niedrigschwellige Methode der logistischen Regression, die bei charakteristikorientierten Ansätzen meist genutzt wird, dazu eignet.

Anmerkung

Interessenkonflikte

Die Autoren erklären, dass sie keine Interessenkonflikte in Zusammenhang mit diesem Artikel haben.

Literatur

1. GEK Gmünder Ersatzkasse, Hrsg. GEK-Gesundheitsreport 2003: Auswertungen der GEK-Gesundheitsberichterstattung. St. Augustin: Asgard-Verlag; 2003.
2. Icks A, Rathmann W, Rosenbauer J, Giani G. Robert Koch-Institut, Hrsg. Diabetes mellitus. Berlin: Robert Koch-Institut; 2005. (Gesundheitsberichterstattung des Bundes; 24). URN: urn:nbn:de:0257-1002032
3. Bortz J, Döring N. Forschungsmethoden und Evaluation – für Human- und Sozialwissenschaftler. 4. Aufl. Heidelberg: Springer Medizin Verlag; 2006.
4. Al-Kandari A, Al-Assomi F, Al-Saqabi A, El-Shazly M. Frequent attenders at a primary health care center in Kuwait. Kuwait Med J. 2008;40(1):18-24.
5. Ash AS, Zhao Y, Ellis RP, Schlein Kramer M. Finding future high-cost cases: comparing prior cost versus diagnosis-based methods. Health Serv Res. 2001 Dec;36(6 Pt 2):194-206.
6. Botica MV, Kovacic L, Katic M, Tiljak H, Renar IP, Botica I. Chronic patients--persons with diabetes frequent attenders in Croatian family practice. Coll Antropol. 2007 Jun;31(2):509-16.
7. Etemad LR, McCollam PL. Predictors of high-cost managed care patients with acute coronary syndrome. Curr Med Res Opin. 2005 Dec;21(12):1977-84. DOI: 10.1185/030079905X74970
8. Fleishman JA, Cohen JW. Using information on clinical conditions to predict high-cost patients. Health Serv Res. 2010 Apr;45(2):532-52. DOI: 10.1111/j.1475-6773.2009.01080.x

9. Gregori D, Petrinco M, Barbati G, Bo S, Desideri A, Zanetti R, Merletti F, Pagano E. Extreme regression models for characterizing high-cost patients. J Eval Clin Pract. 2009 Feb;15(1):164-71. DOI: 10.1111/j.1365-2753.2008.00976.x

10. Kapur N, Hunt I, Lunt M, McBeth J, Creed F, Macfarlane G. Psychosocial and illness related predictors of consultation rates in primary care--a cohort study. Psychol Med. 2004 May;34(4):719-28. DOI: 10.1017/S0033291703001223

11. Kersnik J, Svab I, Vegnuti M. Frequent attenders in general practice: quality of life, patient satisfaction, use of medical services and GP characteristics. Scand J Prim Health Care. 2001 Sep;19(3):174-7. DOI: 10.1080/028134301316982405

12. Kobus LSG, Enembreck F, Scalabrin EE, da Silva Dias J, da Silva SH. Automatic knowledge discovery and case management: An effective way to use databases to enhance health care management. IFIP Advances in Information and Communication Technology. 2009;296/2009:241-7. DOI: 10.1007/978-1-4419-0221-4_29

13. Little P, Somerville J, Williamson I, Warner G, Moore M, Wiles R, George S, Smith A, Peveler R. Psychosocial, lifestyle, and health status variables in predicting high attendance among adults. Br J Gen Pract. 2001 Dec;51(473):987-94.

14. Moturu ST, Liu H, Johnson WG. Understanding the effects of sampling on healthcare risk modeling for the prediction of future high-cost patients. In: Fred A, Filipe J, Gamboa H, eds. Biomedical Engineering Systems and Technologies. Berlin, Heidelberg: Springer; 2009. (Communications in Computer and Information Science; 25) p. 493-506. DOI: 10.1007/978-3-540-92219-3_37

15. Moturu ST, Liu H and Johnson WG. Healthcare risk modeling for medicaid patients the impact of sampling on the prediction of high-cost patients. Healthcare risk modeling for medicaid patients the impact of sampling on the prediction of high-cost patients. 2008. In: HEALTHINF 2008 - International Conference on Health Informatics. p. 126-33. Available from: http://www.public.asu.edu/~huanliu/papers/healthInfo08.pdf

16. Rakovski CC, Rosen AK, Wang F, Berlowitz DR. Predicting elderly at risk of increased future healthcare use: How much does diagnostic information add to prior utilization? Health Serv Outcomes Res Methodol. 2002;3(3-4):267-77. DOI: 10.1023/A:1025866331616

17. Roick C, Heider D, Kilian R, Matschinger H, Toumi M, Angermeyer MC. Factors contributing to frequent use of psychiatric inpatient services by schizophrenia patients. Soc Psychiatry Psychiatr Epidemiol. 2004 Sep;39(9):744-51. DOI: 10.1007/s00127-004-0807-8

18. Spießl H, Hübner-Liebermann B, Binder H, Cording C. "Heavy Users" in einer psychiatrischen Klinik - Eine Kohortenstudie mit 1811 Patienten uber fünf Jahre [Heavy users in a psychiatric hospital--a cohort study on 1811 patients over five years]. Psychiatr Prax. 2002 Oct;29(7):350-4. DOI: 10.1055/s-2002-34659

19. Vedsted P, Fink P, Olesen F, Munk-Jørgensen P. Psychological distress as a predictor of frequent attendance in family practice: a cohort study. Psychosomatics. 2001 Sep-Oct;42(5):416-22. DOI: 10.1176/appi.psy.42.5.416

20. Yarger S, Rascati K, Lawson K, Barner J, Leslie R. Analysis of predictive value of four risk models in Medicaid recipients with chronic obstructive pulmonary disease in Texas. Clin Ther. 2008;30 Spec No:1051-7. DOI: 10.1016/j.clinthera.2008.06.001

21. Zhao Y, Ash AS, Haughton J, McMillan B. Identifying future high-cost cases through predictive modeling. Dis Manag Health Outcomes. 2003;11(6):389-97. DOI: 10.2165/00115677-200311060-00005

22. Cucciare MA, O'Donohue W. Predicting future healthcare costs: how well does risk-adjustment work? J Health Organ Manag. 2006;20(2-3):150-62. DOI: 10.1108/14777260610661547

23. LeBlanc M, Moon J, Kooperberg C. Extreme regression. Biostatistics. 2006 Jan;7(1):71-84. DOI: 10.1093/biostatistics/kxi041

24. Backhaus K, Erichson B, Plinke W, Weiber R. Multivariate Analysemethoden: Eine anwendungsorientierte Einführung. 12. vollständig überarbeitete Aufl. Berlin, Heidelberg: Springer; 2009.

25. Ash AS, Ellis RP, Pope GC, Ayanian JZ, Bates DW, Burstin H, Iezzoni LI, MacKay E, Yu W. Using diagnoses to describe populations and predict costs. Health Care Financ Rev. 2000;21(3):7-28.

Korrespondenzadresse:
Justyna Hartmann
Medizinische Hochschule Hannover, Institut für Epidemiologie, Sozialmedizin und Gesundheitssystemforschung, OE 5410, Carl-Neuberg-Str. 1, 30625 Hannover, Deutschland
Hartmann.Justyna@mh-hannover.de

Bitte zitieren als
Hartmann J, Schauer S, Krauth C, Amelung V. Methoden zur Prädiktion von Hochnutzern: ein systematischer Literatur-Review. GMS Med Inform Biom Epidemiol. 2012;8(1):Doc02.
DOI: 10.3205/mibe000126, URN: urn:nbn:de:0183-mibe0001261

Artikel online frei zugänglich unter
http://www.egms.de/en/journals/mibe/2012-8/mibe000126.shtml

Veröffentlicht: *26.09.2012*