DAV

DEUTSCHE
AKTUARVEREINIGUNG e.V.

Ergebnisbericht des Ausschusses Krankenversicherung

Nutzung von Big Data in der Krankenversicherung

Köln, 12.10.2017

Präambel

Der Ausschuss Krankenversicherung der Deutschen Aktuarvereinigung e. V. hat den vorliegenden Ergebnisbericht erstellt.[1]

Zusammenfassung

Der Ergebnisbericht behandelt Fragestellungen zur Nutzung von Big Data in der Krankenversicherung.

Der sachliche Anwendungsbereich dieser Ausarbeitung betrifft die Nutzung von Big Data und Data Analytics bei der Produkt- und Preisgestaltung, Annahmepolitik, Überschussverwendung sowie anderer Anwendungen im weiteren Kontext der Krankenversicherung (z.B. bei Bestandsanalysen, im Leistungsmanagement oder Gesundheitsmanagement, bei der Unternehmensplanung usw.). Des Weiteren werden Fragestellungen der Datengrundlage sowie Datenaufbereitung und statistische Methoden im Zusammenhang mit Big Data behandelt. Schließlich wird abschließend auf Aspekte des Datenschutzes eingegangen. Bei den im Bericht dargestellten Ergebnissen handelt es sich um Zwischenergebnisse der Arbeitsgruppe Big Data. Die Arbeitsgruppe plant, ihre Arbeit mit dem Fokus auf konkrete Berechnungen fortzusetzen und die Ergebnisse in einem weiteren Bericht zu veröffentlichen.

Der Ergebnisbericht ist an die Mitglieder und Gremien der DAV zur Information über den Stand der Diskussion und die erzielten Erkenntnisse gerichtet und stellt keine berufsständisch legitimierte Position der DAV dar.[2]

Verabschiedung

Der Ergebnisbericht ist durch den Ausschuss Krankenversicherung am 12.10.2017 verabschiedet worden.

---

[1] Der Ausschuss dankt der Arbeitsgruppe Big Data ausdrücklich für die geleistete Arbeit, namentlich Daniela Rode (Leitung), Annabritta Biederbick, Dr. Fabian Bohnert, Dr. Jan Esser, Dr. Werner Goldmann, Sebastian Hartmann, Wolfgang Hornung (ausgeschieden), Alexander Jost, Friedrich Loser, Oliver Plahr, David Richter, Friederike Siebert, Dr. Martin Vielitz-Sumi.

[2] Die sachgemäße Anwendung des Ergebnisberichts erfordert aktuarielle Fachkenntnisse. Dieser Ergebnisbericht stellt deshalb keinen Ersatz für entsprechende professionelle aktuarielle Dienstleistungen dar. Aktuarielle Entscheidungen mit Auswirkungen auf persönliche Vorsorge und Absicherung, Kapitalanlage oder geschäftliche Aktivitäten sollten ausschließlich auf Basis der Beurteilung eine(n) qualifizierte(n) Aktuar DAV/Aktuarin DAV getroffen werden.

Inhalt

1. Big Data in der Krankenversicherung

Mit Big Data werden üblicherweise große Datenmengen bezeichnet, welche:

- ein großes Datenvolumen umfassen,
- viele verschiedene Datentypen enthalten,
- kaum oder nicht strukturiert sind sowie
- mit hoher Frequenz auftreten,

so dass sie mit traditionellen Methoden der Datenverarbeitung kaum auszuwerten sind[3]. Im Zusammenhang mit Big Data werden dementsprechend oftmals neue Methoden der Datenauswertung unter Begriffen wie z.B. Data Analytics, Advanced Analytics, Predictive Modelling oder auch Maschinelles Lernen diskutiert.

Die Krankenversicherungssparte – insbesondere die gesetzliche Krankenversicherung sowie die substitutive Versicherung der PKV – weist eine recht hohe Schadenfrequenz auf, da viele Versicherte mehrfach im Jahr einen Arzt aufsuchen oder Behandlungen / Medikamente erhalten. Im Verlauf mehrerer Jahre entstehen durch diese Versicherungsleistungen sehr große Datenbestände. Trotzdem würden diese Datenbestände wohl nicht der obigen Definition von Big Data entsprechen. Gesundheitsdaten in großem Umfang und hoher Frequenz entstehen aber z.B. bei der Nutzung von Fitness- oder Gesundheitstrackern. Die dabei gesammelten Daten zum Gesundheitszustand sowie ggf. Medikation und Therapie könnten sicherlich eher als Big Data angesehen werden.

Dieses Papier befasst sich nicht damit, ob und in welcher Form Big Data in der Krankenversicherung oder angrenzenden Bereichen vorliegt, sondern diskutiert verschiedene Methoden der Advanced Analytics und mögliche Use-Cases im Zusammenhang mit der Krankenversicherungssparte. Es werden die für den Krankenversicherungsbereich üblicherweise erhobenen Daten erläutert sowie ggf. Hinweise auf weitere nützliche Daten gegeben.

Es ist jedoch zu berücksichtigen, dass bei jeder Verarbeitung personenbezogener Daten grundsätzlich das Bundesdatenschutzgesetz zu beachten ist und dass Gesundheitsdaten eine besondere Art personenbezogener Daten darstellen, für die Sonderregelungen gelten. Das Erheben, Verarbeiten und Nutzen von personenbezogenen Daten und vor allem von Gesundheitsdaten ist nur dann zulässig, wenn der Betroffene einwilligt oder ein gesetzlich definierter Ausnahmefall vorliegt. Dies ist individuell je

[3] Vergleiche auch: https://de.wikipedia.org/wiki/Big_Data

nach Ziel und Zweck der Datenauswertung und Art der Anwendung vorab zu prüfen. (Weiterführende Anmerkungen zum Datenschutz sind im Anhang dieses Papiers zu finden.)

2. Nutzung von Big Data bei der aktuariellen Produkt- und Preisgestaltung, Annahmepolitik und Überschussverwendung

Können zusätzlichen Daten, die im Zusammenhang mit Big Data generiert werden, auch bei aktuariellen Fragestellungen der Produkt- und Preisgestaltung in der Krankenversicherung, der Annahmepolitik und Überschussverwendung berücksichtigt werden?

Neben den aufsichtsrechtlichen Rahmenbedingungen (z.B. VAG, VVG, KVAV) sind speziell im Bereich der privaten Vollversicherung ethische Aspekte zu bedenken (Widerspruch zur Eigenschaft der PKV als zweite Säule des Versicherungssystems nach der Einführung der Pflicht zur Versicherung). Die private Vollversicherung wird generell nach Art der Lebensversicherung kalkuliert, wohingegen Zusatzversicherungen sowohl nach Art der Lebensversicherung als auch nach Art der Schadenversicherung kalkuliert werden können. Für Tarife, die nach Art der Lebensversicherung betrieben werden, ist der aufsichtsrechtliche Rahmen eng gefasst: Die Gewährung von Beitragsvorteilen innerhalb eines Tarifs durch eine Reduktion der Kopfschäden führt zu getrennten Kollektiven und damit getrennten Tarifen. Pauschale, altersunabhängige Zuschläge sind in der Regel nur bei Vorerkrankungen möglich. Vor dem Hintergrund des Gleichbehandlungsgrundsatzes könnten auch Preisdifferenzierungen über differenzierte Kostenansätze oder einen differenzierten Sicherheitszuschlag (z.B. mangels Begründbarkeit) unzulässig sein. Selbst bei ausreichender Datengrundlage und Begründbarkeit einer preislichen Differenzierung sind aber nachteilige Bewertungen (z.B. bei Risikoverschlechterungen) nach Vertragsabschluss rechtlich unzulässig. Eine nachhaltige Kalkulation scheint damit fraglich.

Bei Zusatzversicherungen bietet die Kalkulation nach Art der Schadenversicherung hier generell mehr Freiheitsgrade, insbesondere durch die Möglichkeiten zur unterschiedlichen Ausgestaltung des Kündigungsrechts bzw. zur Befristung. Bei einer grundsätzlich denkbaren individualisierten Kalkulation z.B. über Scoring-Modelle (analog z.B. Kfz-Versicherung) scheint ein Kündigungsrecht des Versicherers bzw. eine Befristung sogar zwingend erforderlich.

Neben der preislichen Differenzierung kommt im Rahmen der Produktgestaltung außerdem eine Leistungsdifferenzierung in Betracht. Bei der Ausgestaltung des Tarifdesigns ist ein Bezug zu typischen Krankenversicherungsleistungen erforderlich. Eine pauschale Leistung beispielsweise allein wegen der Nutzung einer App dürfte dem Verbot versicherungsfremden Geschäfts widersprechen. Leistungsdifferenzierungen wie zum Beispiel über das Bonusheft der Zahnabsicherung in der GKV sind aber denkbar.

Bei der aktuariellen Betrachtung der Annahmeregeln inkl. der Risikoprüfung muss zwischen Neugeschäft und Bestand unterschieden werden. Aufgrund der Vertragsfreiheit könnten im Neugeschäft durchaus Erkenntnisse aus erweiterten Daten zur Ablehnung eines Antrags führen. Wird ein Antrag aber angenommen, sind Risikozuschläge bzw. Leistungsausschlüsse bei Tarifen nach Art der Lebensversicherung in der Regel nur wegen Vorerkrankungen möglich. Im Bestand ist zudem das Tarifwechselrecht zu beachten, sofern die Versicherungsfähigkeit im Zieltarif gegeben ist; darüber hinaus können Risikozuschläge bzw. Leistungsausschlüsse nur auf Mehrleistungen erhoben werden. Auch ist fraglich, ob die Versicherungsfähigkeit (im Ausgangs- oder im Zieltarif) an die Einwilligung in die Erhebung von bestimmten Daten geknüpft werden kann.

Bei der Verwendung von Überschüssen aus der Rückstellung für Beitragsrückerstattung ist ebenfalls eine Berücksichtigung weitergehender Daten denkbar. Unter dem Gesichtspunkt der Gleichbehandlung ist eine differenzierte Mittelvergabe aber rechtlich nur dann zulässig, wenn sie versicherungstechnisch wesentlich ist. Zum Beispiel scheint eine BRE-Auszahlung allein gekoppelt an die Bereitschaft zur App-Nutzung fraglich. Andererseits ist eine Differenzierung nach gesundheitsbewusstem Verhalten denkbar.

Fazit: Aufgrund der regulatorischen Vorgaben insbesondere durch die KVAV werden in der nach Art der Lebensversicherung kalkulierten Voll- und Zusatzversicherung derzeit kaum Ansätze für eine preisliche Differenzierung durch die Nutzung zusätzlicher Daten gesehen. Mehr Möglichkeiten bietet die Kalkulation nach Art der Schadenversicherung. Über die dargestellten Produktthemen hinaus dürften die zunehmenden Daten und Informationen im Zusammenhang mit der Digitalisierung aber weitere Handlungsfelder eröffnen. Dies wird in den nachfolgenden Abschnitten betrachtet.

3. Diskussion weiterer Anwendungsgebiete von Big Data in der Krankenversicherung

Big Data könnte jedoch bei verschiedenen Use-Cases der Krankenversicherung unterstützend verwendet werden. Bei allen Anwendungsfällen und generell bei Big Data-Analysen sind grundsätzlich die datenschutzrechtlichen Regelungen zu beachten. (Weiterführende Anmerkungen zum Datenschutz sind im Anhang dieses Papiers zu finden.) Dies gilt in der Regel bereits zu dem Zeitpunkt, bevor eine Analyse durchgeführt wird, um sicherzustellen, dass die Analyse selbst datenschutzrechtlich unbedenklich ist.

Folgende Themenbereiche der Krankenversicherung bieten sich für Datenanalysen besonders an:

- Leistungsdatenanalyse / Leistungsprognose,
- Gesundheitsmanagement,

- Bestandsanalysen und Angebotserstellung,
- Prozessautomatisierung / Dunkelverarbeitung.

Mithilfe von Big Data und Data Analytics könnte z.B. die Betrugserkennung im Leistungsbereich oder bei der Verletzung vorvertraglicher Anzeigepflichten verbessert werden.

Durch die personenbezogene Analyse von Krankheitsverläufen könnte das Leistungsmanagement optimiert werden. Dazu wären Maßnahmen mit positiver Wirkung im Zeitablauf zu identifizieren und das Einsparpotential sowie der Kundennutzen abzuschätzen. Durch ein entsprechendes Fallmanagement könnten die Betroffenen identifiziert, deren aktuelle Behandlung verbessert bzw. angepasst und das Auftreten von Folgeerkrankungen vermieden oder herausgezögert werden.

Die Vorhersage der künftigen Erkrankungen und Leistungsbeträge für jeden einzelnen Versicherten könnte insgesamt zu genaueren Leistungsprognosen und damit zu einer verbesserten Einschätzung der zukünftigen Risikolage des Versicherers führen.

Big Data könnte auch Präventionsmaßnahmen unterstützen. Denkbar wäre hier zum Beispiel die Erfassung und Bewertung verschiedener Gesundheitsdaten, sportlicher Aktivitäten und der Ernährung. Der Kunde könnte speziell auf seine Gesundheitssituation abgestimmte Hinweise, Erinnerungen an Termine oder Medikamente bzw. Belohnung bei Zielerreichung erhalten. Darüber hinaus könnte das Controlling von Gesundheitsmanagementmaßnahmen optimiert werden.

Ein weiterer Anwendungsbereich von Big Data und Data Analytics liegt in der Bestandsanalyse. Beispielhaft könnten durch Scoringverfahren das individuelle oder kollektive Stornoverhalten genauer prognostiziert und ggfs. Stornovermeidungsmaßnahmen initiiert werden. Außerdem könnten Versicherungsangebote passgenauer zugeschnitten, affine Kunden identifiziert und verstärkt als gut einzuschätzende Risikogruppen kontaktiert werden. Eine Analyse der Preissensitivitäten der Versicherten wäre vor allem im Bereich der nach Art der Schadenversicherung kalkulierten KV denkbar.

Eine verstärkte Automatisierung wäre sowohl im Leistungsbereich als auch bei der Gesundheitsprüfung / Angebotserstellung möglich. So könnte zum Beispiel eine automatisierte Analyse der eingereichten Rechnung (Dunkelverarbeitung) im Gesamtkontext der Leistungshistorie des jeweiligen Versicherten und im Vergleich zu typischen Krankheitsverläufen, Krankheitskosten, Behandlungen / Therapien erfolgen. Außerdem wären mittels Daten zu bestimmten Krankheitsbildern, Verhaltensmerkmalen (Raucher, Sport, gesundheitsbewusste Ernährung, Labordaten), Bildungsgrad, Woh-

nort usw. und einem entsprechenden Abgleich mit Erfahrungen aus dem Bestand automatisierte Gesundheitsprüfungen / Angebotserstellungen denkbar. Eine "bildunterstütze" Gesundheitsprüfung wäre ebenfalls vorstellbar.

Im Folgenden werden beispielhaft drei Use-Cases detaillierter beschrieben:

a. Use-Case „Metrische Leistungsschätzung“

Dieser Use-Case beschäftigt sich mit der Vorhersage der zu erwartenden Versicherungsleistungen einzelner versicherter Personen in einem bestimmten Zeitraum – typischerweise in einem Kalenderjahr. Diese Kenntnis kann zu unterschiedlichen Zwecken verwandt werden:

- Durch Summation über Teilmengen des Versichertenbestands können Vorhersagen der Leistungsentwicklung dieser Teilbestände gewonnen werden.
- Es können künftige Hochkostenfälle ermittelt werden.
- Risiken, die sich günstig entwickeln, können bestimmt werden.

Die so gewonnenen Erkenntnisse können einerseits in der Unternehmensplanung einfließen, andererseits zum konkreten Bestandsmanagement eingesetzt werden. Mögliche Anwendungen sind u.a.:

- **Angebotsaktionen**
  Wenn vorhergesagt werden kann, für welche versicherten Personen sich die Leistungsausgaben voraussichtlich günstig entwickeln, können hier gezielt Upsellingangebote erstellt werden.
- **Leistungsmanagement (Case-/Disease-Management)**
  Die Kenntnis voraussichtlicher künftiger Hochleistungsfälle kann genutzt werden, die Leistungsentwicklung dieser Personen gezielt zu analysieren, um zu prüfen, ob durch prophylaktische Maßnahmen Krankheiten verhindert oder zumindest abgemildert werden können – und damit ggf. auch Leistungsausgaben verhindert oder vermindert werden können.
- **Kalkulation**
  Durch die Prognose der Leistungen jeder einzelnen Person können ggf. zusätzliche Daten berücksichtigt werden, die bei einer Prognose auf Basis verdichteter Daten nicht einfließen. So besteht die Möglichkeit, durch Aggregation der prognostizierten Leistungen aller Versicherten einer Beobachtungseinheit eine bessere Vorhersage der kurz-oder mittelfristig zu erwartenden Leistungen zu erhalten als dies bei klassischen Methoden basierend auf verdichteten Daten möglich ist. Außerdem kann man eine Streuung der Daten berechnen und eine Schadenhöhenverteilung herleiten und so z.B. die Quantile dieser Verteilung bestimmen. Damit kann die Sicherheit der Kalkulation überprüft werden.

- **Rückversicherung**
  Die Kenntnis der erwarteten Verteilung künftiger Leistungshöhen kann bei der Entscheidung genutzt werden, ob bzw. welche Rückversicherungen abgeschlossen werden sollten.
- **Unternehmensplanung / RfB-Steuerung**
  Die Bestimmung einer Schadenhöhenverteilung kann insbesondere bei der Unternehmensplanung wichtig sein. Soll eine Beitragsanpassung zu Lasten der RfB auf eine bestimmte Höhe begrenzt werden, so ist es entscheidend, zu wissen, für wie viele Versicherte die unlimitierte Beitragserhöhung wie hoch über dieser Grenze liegt, um den benötigten Einmalbeitrag zu ermitteln. Um über mehrere Jahre zu planen, welche RfB-Mittel für Limitierungen der Beitragsanpassungen jeweils eingesetzt werden sollen, ist die Kenntnis der Schadenhöhenverteilung notwendig.
- **Risikoprüfung**
  Soweit es gelingt nicht nur kurzfristige Prognosen der künftigen Leistungen, sondern auch lang- oder zumindest mittelfristige Vorhersagen zumachen, kann ein solches Verfahren auch zu Risikoprüfung eingesetzt werden. Voraussetzung hierzu ist allerdings ein Verfahren, dass eine funktionale Abhängigkeit der erwarteten Leistungen von den Personendaten liefert.

Insbesondere unter dem Gesichtspunkt möglichst langfristige Prognosen herzuleiten, sollte auch der Auswertungszeitraum „viele“ Jahre umfassen, wobei selbstverständlich eine Aufbereitung notwendig ist, um z.B. Systembrüche wie eine GOÄ-Änderung oder eine Änderung der Definition des Pflegebegriffs zu berücksichtigen. Ob und wie weit eine solche Datenaufbereitung sinnvoll ist, ist jeweils nur im Einzelfall zu entscheiden.

Der Nutzen für die Versicherungsunternehmen liegt je nach Anwendung z.B. in der Verbesserung der Planung, einem höheren Bewusstsein über die Höhe des Risikos bei der Kalkulation, besserer Steuerung des Leistungsmanagements usw. Aber auch aus Kundensicht ergeben sich Vorteile durch verbesserte Möglichkeiten der Beratung durch den Versicherer. Auch eine verbesserte Unternehmenssteuerung und RfB-Planung wirken sich in Form höherer Überschüsse und damit erhöhter RfB-Zuführung positiv auf die Versichertengemeinschaft aus.

b. Use-Case "Vorhersage von Ereignissen"

Dieser Use-Case beschäftigt sich mit der Vorhersage des Eintritts eines festgelegten Ereignisses wie z.B. einer bestimmten Erkrankung oder einem Krankenhaus-Aufenthalt (KH-Aufenthalt) für eine einzelne Person. Mit dieser Information könnte der Versicherer u.a. Steuerungsmaßnahmen zur Gesunderhaltung, Prävention oder auch zum Diseasemanagement ableiten. Bei einer Reduzierung des Erkrankungsrisikos bzw. bei

Vermeidung oder Verkürzung eines KH-Aufenthalts bzw. bei einer gesteuerten terminierten Einweisung im Vergleich zu einer Notarzt-Einweisung wird von einem Nutzen für den Versicherten aber auch für das Versicherungsunternehmen ausgegangen.

Ein Nutzen kann nur entstehen, wenn die Möglichkeit einer Einflussnahme besteht. D.h. der Versicherte muss Handlungsoptionen und entsprechend Zeit zur Handlung haben. Dies sollte bei der Auswahl des Ereignisses, z.B. der Wahl der vorherzusagenden Erkrankung berücksichtigt werden. Als Vorhersagezeitraum halten wir einen Horizont von 1 - 2 Jahren für angemessen. Einerseits sollte der Zeitraum nicht zu lang sein (z.B. in den nächsten 10 Jahren tritt eine schwere Erkrankung auf / steht eine stationäre Behandlung an) - andererseits sollte er auch nicht zur kurz sein, um eine Reaktion bzw. Beeinflussung / Steuerung des Versicherten zu ermöglichen.

Der Auswertungszeitraum sollte für die stationären Daten mehrere Jahre z.B. 5 Jahre umfassen und für die ambulanten Daten mind. 2 Jahre. Es sollte darauf geachtet werden, dass keine Vermischung von Vergangenheits- und Zukunftsdaten erfolgt. Wenn der Eintritt eines Krankenhausaufenthalts z.B. über das Aufnahmedatum (mit anschließend mind. einer Übernachtung) definiert wird, muss darauf geachtet werden, dass man zur Modellbildung nur die Daten aus dem Gesamtdatensatz benutzt, die bis zu diesem Zeitpunkt vorgelegen haben und die anderen für den Lerndatenbestand entfernt. Ansonsten wird auf falschen Annahmen gelernt und das Modell ist nicht wirksam auf dem Echtdatenbestand.

Des Weiteren müssen Abgrenzungsproblematiken untersucht und gelöst werden. Z.B. bei welcher Diagnose / welcher Kombination von Diagnosen / welchem Schweregrad liegt eine bestimmte Erkrankung vor? Muss diese mindestens für eine gewisse Dauer bestanden haben? Bei Krankenhausaufenthalten soll in diesem Use-Case auf Erkrankungen abgestellt werden, d.h. Unfallereignisse werden nicht betrachtet und sind deshalb auszusteuern. Weiterhin ist es ggf. angeraten, auch Krankenhausaufenthalte zur Entbindung bzw. aufgrund von Schwangerschaft auszusteuern, da diese nicht die Zielgruppe für diesen Use-Case darstellen. Es kann des Weiteren sinnvoll sein, bei der Modellierung zwischen körperlichen und psychischen Erkrankungen zu unterscheiden bzw. diese getrennt zu modellieren, da die Erkrankungsmuster und -zeiträume sehr unterschiedlich sind.

Das Modell in diesem Use-Case soll Versicherte identifizieren, bei denen mit hoher Wahrscheinlichkeit innerhalb von 1 – 2 Jahren eine bestimmte Erkrankung auftritt oder auch eine Behandlung im Krankenhaus notwendig wird. Diese Versicherten könnte das Versicherungsunternehmen innerhalb des o.g. Zeitraums dann kontaktieren, um individuelle Vorschläge zu Handlungsoptionen zu machen. Z.B.:

- Bei noch nicht manifesten Erkrankungen, aber einer hohen Gefährdung: Vorschläge zur Verringerung der zukünftigen Risikosituation → z.B. Vorschläge zum Life-Style / Ernährung / Sport, Präventions-Programme

- Bei Versicherten mit bestehenden Erkrankungen und konkreter Annahme einer Verschlechterung: Vorschläge zur Verbesserung der Risikosituation → z.B. Vorschläge zum Life-Style / Ernährung / Sport, Diseasemanagement-Programme
- Hinweis zur Kontaktierung des Hausarztes / eines Facharztes z.B. für eine Untersuchung bei Verdacht auf eine mögliche schwere Erkrankung (die bisher nicht eingetreten ist / diagnostiziert wurde) bzw. um die Behandlung oder Medikation einer bestehenden Erkrankung zu überprüfen
- Unterstützung bei der Terminierung oder passenden Wahl eines Krankenhauses, wenn ein KH-Aufenthalt notwendig ist bzw. Unterstützung bei der Überbrückung des Zeitraums bis zum KH-Aufenthalt, wenn eine Wartezeit und das Risiko einer Komplikation / Verschlechterung bestehen.

Der Nutzen für den Versicherten liegt in der Unterstützung seiner Gesunderhaltung und der Vermeidung / Verkürzung eines Krankenhausaufenthaltes sowie der Unterstützung im Umfeld eines möglichen Eintritts schwerer Erkrankungen. Der Nutzen des Versicherungsunternehmens kann einerseits in Einsparungen von Leistungskosten bestehen und andererseits in dem Nutzen einer besseren Betreuung des Versicherten und dessen gestiegener Lebensqualität. Dem gegenüber stehen Kosten durch die gesonderte Kontaktierung und Betreuung des Versicherten sowie ggf. zusätzliche Leistungsausgaben durch besondere Therapien oder Ausgaben durch die Einrichtung und Durchführung von speziellen Präventions- oder Diseasemanagement-Programmen.

c. Use-Case „Propensity Score Matching“

„Propensity Score Matching“ kann verwendet werden, um den Nutzen einer Maßnahme, eines Programms oder einer Therapie zu messen. Bei vielen Programmen ist es schwierig, den Nutzen direkt zu ermitteln und die Programme oder Maßnahmen zu evaluieren. Das Ergebnis der Nichtteilnahme der Teilnehmer bzw. der Teilnahme von Nicht-Teilnehmern kann nicht beobachtet werden. Die Vergleichbarkeit von Teilnehmern und Nicht-Teilnehmern ist aber nicht per se gegeben, denn die beiden Gruppen könnten sich systematisch unterscheiden.

Deshalb werden basierend auf mehrdimensionalen statistischen Merkmalen der Einzelpersonen (z.B. Alter, Geschlecht, sozioökonomische Parameter, Leistungsausgaben, Komorbiditäten) Teilnahmewahrscheinlichkeiten geschätzt. Anschließend werden Teilnehmer/Nicht-Teilnehmer-Paare mit (annähernd) gleicher Teilnahmewahrscheinlichkeit gebildet. Anhand dieser Paare wird der Nutzen gemessen.

Anwendungsbeispiele:

- Effekte auf Leistungsausgaben von Programmen bei bestimmten Erkrankungen (z.B. Diabetes)

- Effekte auf Storni bei verschiedenen Provisionsmodellen (z.B. Abschluss- vs. Bestandsprovision)
- Effekte auf Verwaltungsaufwände neuer Kommunikationskanäle (z.B. Rechnungs-App)

Mit den gewonnenen Erkenntnissen können die durchgeführten Aktionen überprüft und gegebenenfalls neu justiert werden.

Der Auswertungszeitraum sollte auch hier „viele“ Jahre umfassen, um langfristige Effekte (z.B. Folgeerkrankungen bei Diabetes) miterfassen und bewerten zu können. Sondereinflüsse sind dabei angemessen zu berücksichtigen, z.B. neuartige Medikamente und Methoden zur Behandlung oder sogar Heilung einer Erkrankung.

Darüber hinaus ist die Frage zu klären, ob monetär basierte Kosten-/Nutzenanalysen zur finalen Beurteilung einer Intervention ausreichen. Indirekte Effekte z.B. auf die Kundenzufriedenheit, die Weiterempfehlungsbereitschaft oder das Image des Versicherungsunternehmens könnten mittels Kundenbefragung oder Erhebung der gewonnenen (gesunden) Lebensjahre ergänzt werden.

4. Datengrundlage in der Krankenversicherung und Hinweise zur Datenaufbereitung

a. Datengrundlage

In der Krankenversicherung werden üblicherweise personenbezogene Daten zur Durchführung des Versicherungsvertrages erhoben und verwendet.

Man kann einerseits unterscheiden zwischen Daten, die Informationen zur Person beschreiben wie z.B. Geschlecht, Alter, Beruf oder Wohnort sowie Angaben zur Art des abgeschlossenen Versicherungsschutzes, z.B. Voll- oder Ergänzungsversicherung, Premium- oder Basisschutz, Höhe des Selbstbehalts oder eine eventuelle BRE.

Daneben liegen verschiedenste Daten vor, die den Gesundheitszustand einer versicherten Person beschreiben. Viele Informationen zu Diagnosen, Medikation und Therapie können den zur Leistungserbringung eingereichten Rechnungen und Belegen entnommen werden. Dies sind insbesondere GOÄ- und GOZ-Ziffern, soweit vorhanden ICD-Codes, DRG oder Operationsschlüssel (OPS), Rezeptdaten oder Informationen über Heil- und Hilfsmittel. Wichtig für die Leistungsabrechnung sind zudem die Behandlungsdaten und Krankheitsdauern, in der Pflegeversicherung auch die Pflegegrade sowie die Höhe des Rechnungsbetrags an sich. Außerdem liegen in der Regel nähere Informationen über die Leistungserbringer vor, zum Beispiel die Fachrichtung des behandelnden Arztes. Demgegenüber gibt es jedoch wichtige Gesundheitsdaten, die üblicherweise beim Krankenversicherer häufig nicht bzw. nicht aktuell oder nicht

umfassend vorliegen. Dazu zählen z.B. der BMI, Raucherstatus, Informationen zu Ernährung oder sportlicher Aktivität, aber auch Blutdruck und Laborwerte sowie deren Volatilität.

Des Weiteren können externe Daten von Interesse sein. Dies können demografische Daten sein, z.B. die regionale Arztdichte oder die Entfernung zum nächstgelegenen Krankenhaus. Soweit sie Krankheiten beeinflussen, können auch Umweltbedingungen wie zum Beispiel die Luftqualität oder die Lärmbelastung von Bedeutung sein.

b. Gruppierung und Verdichtung

Beim Umgang mit großen Datenmengen kommen häufig Methoden der Aggregation zum Einsatz. Dies soll im Folgenden beispielhaft anhand des ICD (International Statistical Classification of Diseases and Related Health Problems) dargestellt werden. Der ICD ist das wichtigste, weltweit anerkannte Diagnoseklassifikationssystem der Medizin und spielt naturgemäß bei der Bewertung und Analyse von Daten im Krankenversicherungsbereich eine wichtige Rolle.

Die Leistungshistorie einer vollversicherten Person kann nach mehreren Jahren eine große Menge von Daten mit einer Vielzahl verschiedener Diagnosen enthalten. Die Modellierung auf dieser Flut von Einzelinformationen über differenzierte und isolierte Einzelleistungen gestaltet sich oft schwierig. Aufgrund der Vielzahl der ICD-Codes (aktuell ca. 16.000) bietet es sich daher oftmals an, vor Modellierung die einzelnen Diagnosen zusammenzufassen. Dabei kann man verschiedene Wege gehen: Z.B. können Diagnosen-Codes rein fachlich gruppiert werden – unabhängig von der individuellen Leistungshistorie der versicherten Person. Eine andere Herangehensweise ist z.B. die Verdichtung der Diagnose-Codes auf Krankheitsepisoden, d.h. eine zeitlich und fachliche Zusammenfassung der Erkrankungshistorie eines Versicherten.

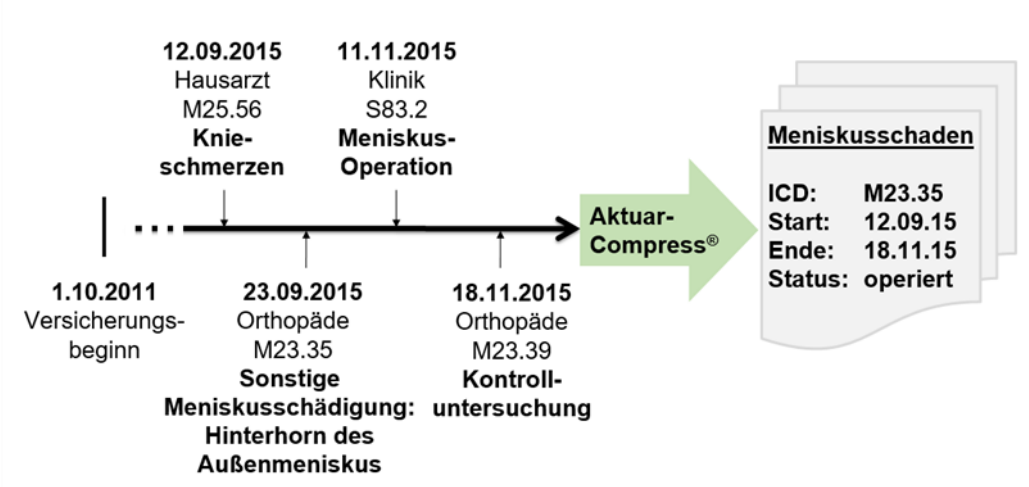
Ein Beispiel für eine fachliche Gruppierung ist das in der GKV benutzte hierarchische Klassifikationsverfahren zur Gruppierung von Erkrankungen[4]. Dabei werden die ICD-Codes nach klinischer Homogenität zu knapp 1.000 diagnosebezogenen Risikogruppen zusammengefasst. Über die Pharmazentralnummer können noch Informationen aus den Medikamenten dazugespielt werden und auf deren Wirkstoffbasis vertiefende Informationen generiert werden (z.B. Schweregrade einer Erkrankung oder ob eine chronische Erkrankung vorliegt) sowie zusätzlich Plausibilisierungen vorgenommen werden.

---

[4] Details zur Klassifikation finden sich z.B. bei der Beschreibung der jährlichen Festlegungen des Bundesversicherungsamtes im Rahmen des morbiditätsorientierten Risikostrukturausgleichs unter: http://www.bundesversicherungsamt.de/risikostrukturausgleich/festlegungen.html

Ein Beispiel für die Verdichtung auf Krankheitsepisoden stellt das Verfahren *Aktuar-Compress*® von RISK-CONSULTING Prof. Dr. Weyer GmbH dar, das Leistungsdaten verdichtet und fokussiert. Alle Abrechnungsdaten einer Erkrankungsepisode einer versicherten Person werden zusammengefasst und auf einen führenden ICD-Code verdichtet. Dazu wurden ca. 700 Krankheitsbilder definiert, die medizinisch und risikotechnisch kohärente Diagnosen beinhalten sowie auch die entsprechenden Co-Diagnosen und Erkrankungsentwicklungen / -stufen. Alle mit einer Erkrankung einhergehenden Diagnosen werden so integriert. Dieses Verfahren ermöglicht auch die Plausibilisierung der zugrundeliegenden Daten anhand verfügbarer Sekundärinformationen aus den Leistungsdaten – u.a. Kosten, Dauer, Häufigkeit sowie Art der Behandlung etc. Auch unklare oder ungenaue Diagnosen wie zum Beispiel "Rückenschmerzen" können über den Kontext aufgelöst werden – der Schlüssel hierzu ist die Betrachtung der gesamten Leistungshistorie, nicht nur der einzelnen Abrechnung.

Beispiel einer Verdichtung



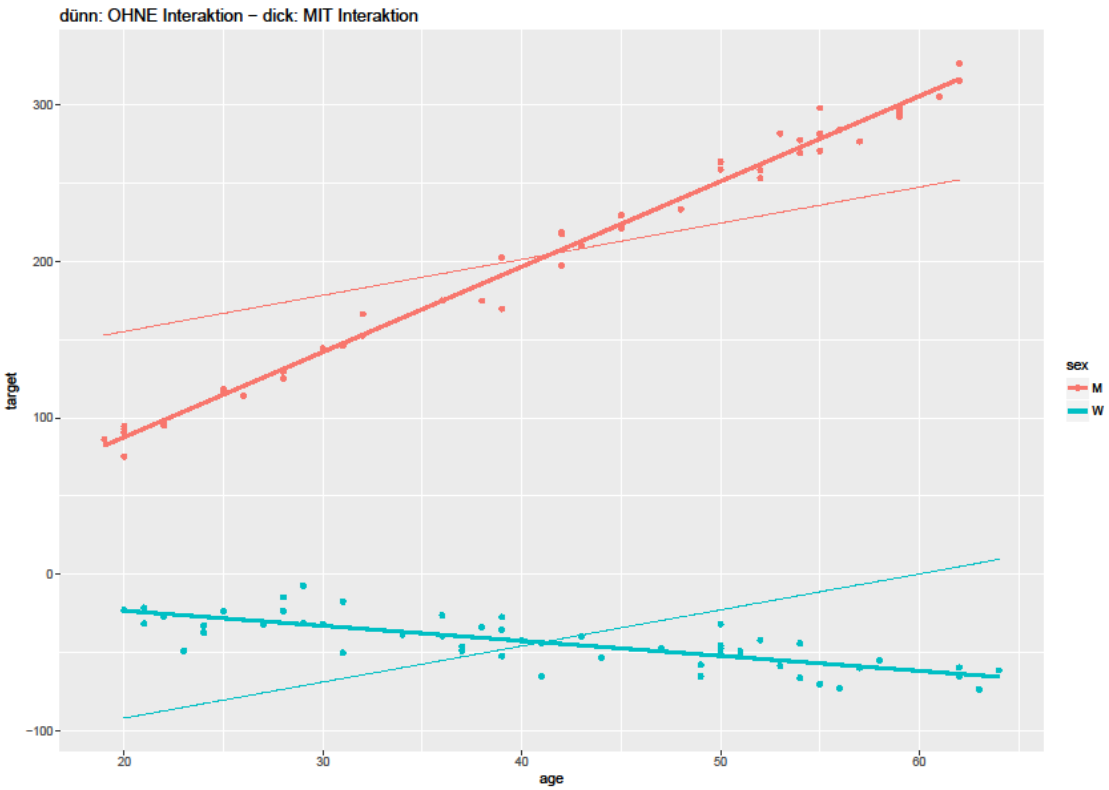
c. Datenaufbereitung

Die Datenaufbereitung spielt eine entscheidende Rolle in jeder Analyse. Viele Modelle können nur bestimmte Datenausprägungen (stetig oder kategorisch) gut verarbeiten. Manche Modelle reagieren empfindlich gegenüber Ausreißern bei den Prädiktoren, andere nicht so sehr. (Ausreißer bei der Zielvariablen müssen stets behandelt werden).

Ziel der Datenaufbereitung ist es daher, den Datensatz optimal für die verwendeten Modelle aufzubereiten. Darüber hinaus ist es häufig sinnvoll, den vorhandenen Datensatz um weitere Merkmale zu erweitern. Zu beachten ist, dass die Auswahl und die Aufbereitung der Daten die späteren Ergebnisse der Modelle beeinflussen. Fehler die an dieser Stelle gemacht werden, wirken sich dementsprechende auf das Resultat der Analyse aus.

Folgende Schritte zur Datenaufbereitung sind daher angeraten:

- **Weitere Merkmale** ermitteln: Einerseits können direkte Ableitungen aus einem Merkmal erforderlich sein (z.B. Alter aus Geburtsdatum) – andererseits können durch Kombinationen von Merkmalen neue wichtige Einflussgrößen ermittelt werden (z.B. Kombination von Alter der versicherten Person mit klinischen Werten, ICD-Codes, DRG-Codes; Entfernung zum nächsten Krankenhaus aus Adresse des Krankenhauses und Wohnort des Versicherten). Gegebenenfalls entstehen hier sehr viele Merkmale, wodurch Laufzeiten erhöht werden und unter Umständen die Stabilität und Treffsicherheit des Modells negativ beeinflusst wird. In solchen Fällen können vorab Verfahren zur Merkmalsreduktion angewandt werden (z.B. die Hauptkomponentenanalyse, auch PCA genannt).
- **Wechselwirkungen** modellieren: Je nach Modell ist es erforderlich, Interaktionen zwischen Variablen separat zu modellieren. Beispielsweise kann der kombinierte Einfluss von Alter und Geschlecht in einem linearen Modell nicht immer adäquat abgebildet werden. Dies sehen wir in der folgenden Abbildung: Die Variable *target* wird in Abhängigkeit von den beiden Variablen Alter *age* und Geschlecht *sex* linear modelliert. Die dünnen Linien zeigen die Vorhersagen eines linearen Modells ohne Interaktionsterm (Prädiktoren: *age*, *sex*). Die dicken Linien zeigen die Vorhersage eines linearen Modells, in das ein Interaktionsterm "Alter mal Geschlecht" als neue Variable eingeführt worden ist (Prädiktoren: *age*, *sex*, *age*sex*).



- **Fehlerhafte Daten** müssen identifiziert und bereinigt werden: Z.B. fehlerhafte ICD-Codes aus Scanfehlern → Codes entweder löschen oder korrigieren. Man kann an diese Stelle auch schon statistische Verfahren nutzen, um fehlerhafte Daten zu identifizieren (z.B. Assoziationsanalyse, Clusteranalyse).
- Falls das statistische Verfahren metrische Variablen voraussetzt (z.B. Regressionsverfahren), müssen **nominale und ordinale Variablen** umcodiert werden. Eine gängige Herangehensweise ist das Dichotomisieren, in dem alle möglichen Ausprägungen einer nominalen Variable in entsprechend viele "Ja / Nein"-Felder zerlegt wird. (Z.B. das Merkmal Berufsgruppe mit 10 Ausprägungen wird durch das Dichotomisieren in zehn Merkmale, d.h. in 9 Berufsgruppen[5] mit jeweils der Ausprägung 0 / 1 umgewandelt.) Eine andere Möglichkeit ist das "Indizieren", bei dem den einzelnen Ausprägungen der nominalen Variable im Hinblick auf die Zielvariable "sinnvolle / vernünftige" Zahlenwerte zugeordnet werden. Hierbei sieht man sofort, dass die Schwierigkeit bei dieser Variante die Bestimmung der "sinnvollen" Werte ist; diese muss zur späteren Modellierung passen und ggf. zyklisch angepasst werden. (Gleichzeitig ist jedoch Overfitting zu vermeiden…)
- Manche Modelle nehmen die statistische **Unabhängigkeit oder Unkorreliertheit der Prädiktor-Variablen** an, damit beweisbare Eigenschaften der Schätzer abgeleitet werden können. Auf die exakte Erfüllung dieser Bedingungen kann in einer bewussten Entscheidung verzichtet werden. Jedoch profitieren solche Modelle in der Regel davon, wenn möglichst wenig Korrelation zwischen den Prädiktor-Variablen besteht. Um dies zu erreichen bietet sich das Weglassen von Prädiktor-Variablen oder etwa die Hauptkomponentenanalyse (PCA), welche die Prädiktoren geeignet linear transformiert, an.
- Falls das statistische Verfahren metrische Variablen voraussetzt (z.B. Regressionsverfahren), müssen auch **missing values** speziell behandelt werden. Die einfachste Möglichkeit ist, Datensätze mit missing values zu eliminieren, d.h. aus der Analyse zu entfernen. Hierbei besteht jedoch die Gefahr einen systematischen Fehler zu begehen (wenn die missig values nicht zufällig verteilt sind) oder die Trainingsdaten zu stark zu reduzieren und somit Aussagekraft zu verlieren (da der Datensatz immer komplett eliminiert werden muss). Eine andere Möglichkeit mit missing values umzugehen ist, missing values einen

[5] Dass aus einem kategorialen Merkmal mit n Ausprägungen (n-1) dichotome Merkmale (mit jeweils den beiden Ausprägungen "ja, diese Kategorie" und "nein, nicht diese Kategorie" kodiert werden, ist technischer Natur.

"sinnvollen" numerischen Wert zuzuordnen, sogenanntes "Imputing". Beispielsweise kann der Mittelwert der nicht-missing Ausprägungen zugeordnet werden).

- Insbesondere bei spärlich besetzten Datenmatrizen (**sparse matrix**), kann es sinnvoll sein, den benötigten Speicherplatz zu reduzieren. Dazu gibt es verschiedene Vorgehensweisen, auch in Abhängigkeit von der Modellierung. Ein einfaches Beispiel ist die so genannte "Faktentabelle", wo nur die in der Matrix befüllten Zellen abgelegt werden. Die meisten Statistikumgebungen halten spezielle Datentypen und Algorithmen für dünn besetzte Matrizen vor.
- Von praktischer Relevanz ist es, die Aufbereitung der Daten in einen weitgehend **automatisierten Prozess** zu verlagern, der mit einer Versionsnummer versehen ist. Somit kann sichergestellt werden, dass stets nachvollziehbar ist, wie aus den Rohdaten der Input für ein spezielles Modell entstanden ist.

5. Überblick zu statistischen Methoden in der Analyse von Big Data

Das Ziel des folgenden Kapitels ist, einen Überblick über ausgewählte statistische Methoden zu geben. Es erhebt dabei nicht den Anspruch, eine vollständige Auflistung der Methoden bereitzustellen oder zur Vertiefung zu dienen. Vielmehr sollen die verbreiteten Methoden einführend beschrieben und erste Anmerkungen zu Stärken, Schwächen oder Voraussetzungen gemacht werden. Für eine detaillierte Beschäftigung sei auf die Liste einschlägiger Literatur im anliegenden Verzeichnis (Kapitel 0) dieses Ergebnisberichts verwiesen.

a. IT/Voraussetzungen

Die im Folgenden beschriebenen Methoden sind mit einer großen Bandbreite von Anwendungslösungen durchführbar. Sie können einzeln mit quelloffener Software (z.B. Python, R), aber auch mit kommerzieller Data Mining Software (z.B. SAS, SPSS) angewendet werden. In den Mining Suites ist typischerweise die komplette Prozesskette von der Datenaufbereitung über Stichprobenbildung, Modelloptimierung und -bewertung sowie schließlich die Generierung eines Score-Codes für die Modellanwendung umgesetzt. Meist werden dabei gleich mehrere, alternative Methoden miteinander verglichen.

b. Ereignisprognose

Die Vorhersage einer Krankenhauseinweisung, die Vorhersage des Eintritts bestimmter Erkrankungen oder auch die Vorhersage von Storno sind Beispiele von Ereignisprognosen. Vorhersagen zum Eintritt von Erkrankungen sind in der Krankenversicherung insbesondere zur Identifikation von Personen für Gesundheitsmanagementmaßnahmen interessant.

Eine Ereignisprognose kann gut als Klassifikationsaufgabe modelliert werden. Für deren Lösung sind die folgenden statistischen Verfahren grundsätzlich geeignet:

Recht anschaulich, einfach umsetzbar und beliebt ist für diese Fragestellung die Methode **Entscheidungsbaum**. Diese Methode ist für den Einstieg sehr gut geeignet. Die Methoden **„Random Forest“** und **„Tree Boosting“** (Gradient Boosting) bauen auf Entscheidungsbäumen auf und führen i.d.R. zu einer noch besseren Prognose. Bei „Random Forest“ wird mittels „bootstrapping“ und einer Zufallsauswahl der Prädiktormerkmale eine große Anzahl möglichst unterschiedlicher und unabhängiger Entscheidungsbäume gerechnet und damit die Varianz reduziert. „Tree Boosting“ verwendet ebenfalls eine hohe Anzahl an Entscheidungsbäumen. Diese sind jedoch voneinander abhängig und sehr klein. Jeder Baum hängt von den zuvor erzeugten Bäumen ab und versucht die Prognosegenauigkeit zu erhöhen. „Tree Boosting“ gilt als eines der stärksten Verfahren, erfordert aber einige Mühe beim Parametertuning. Dies trifft auch auf **Künstliche Neuronale Netze** zu, die bei Ereignisprognosen derzeit (ggf. noch) nicht so überlegen wie bei der Bild- und Spracherkennung erscheinen. Bei tendenziell linearen Zusammenhängen können mit der **logistischen Regression** gute Prognoseergebnisse erzielt oder eine Benchmark für die rechenintensiveren und unübersichtlicheren Methoden gesetzt werden.

Die folgende Abbildung gibt einen Überblick über einige bei der Modellauswahl für die Ereignisprognose beachtenswerte Kriterien:

Ereignisprognose: **Die Methode ist geeignet für ...**	**Entscheidungsbäume**	**Random Forest**	**Tree Boosting**	**Neuronale Netze**	**Logistische Regression**
fehlende Merkmalswerte	+	+	+	-	-
nichtnumerische Merkmale	+	+	+	-	-
tendeziell knappe Trainingsdaten	+	+	+	-	+
nichtlineare Zusammenhänge	+	+	+	+	-
geringen Tuningbedarf	+	+	-	-	+
geringe Rechenkapazität	+	o	o	-	+
Das Prognosemodell ist ...					
anschaulich und interpretierbar	+	-	-	-	o
überschaubar (geringe Parameteranzahl)	+	-	-	-	+
stabil (z.B bei Datenaktualisierung)	-	-	-	-	+
das mit der besten Vorhersage	-	o	+	+/-	-

Bei der Prognose eines eher seltenen Ereignisses kann durch eine Startstichprobe mit ungefähr gleich vielen Ja/Nein-Datensätzen sowohl eine bessere Beschreibung der Ereignisse als auch ein drastisch geringeres Datenvolumen mit entsprechend kürzeren Laufzeiten erzielt werden.

c. Metrische Prognose

Die Metrische Prognose beschreibt das Problem, für vorgegebene Ausprägungen (bekannter) Eingabevariablen die Werte einer stetigen Zielgröße vorherzusagen. In der Krankenversicherung könnte dies z.B. in der Risikoprüfung genutzt werden, um bei Vorerkrankungen oder anderen vom durchschnittlichen Risiko abweichenden Faktoren den adäquaten Risikozuschlag eines Antragstellers zu ermitteln. Weiterhin könnten mit diesen Verfahren die Leistungskosten versicherter Personen für die Folgejahre geschätzt werden, z.B. für Steuerungsmaßnahmen im Leistungsmanagement oder bei der Unternehmensplanung / Vergabe der RfB-Mittel.

Das mit Abstand bekannteste Verfahren für metrische Prognosen ist die **lineare Regression**. In diesem einfachen Modell wird ein linearer funktionaler Zusammenhang zwischen Eingabe (x) und Zielgröße (y) unterstellt (y= a x), welcher durch normalverteilte Fehlerterme gestört wird.

Eine Erweiterung der linearen Regression bilden die **GLM** (**Generalisierte Lineare Modelle**). Dieser Modellklasse liegt die wesentlich schwächere Annahme eines linearen Zusammenhangs zwischen der Eingabe (x) und der mittels einer monotonen, differenzierbaren Funktion (g) abgebildeten Zielgröße zugrunde (x=g(y)). Zudem können in den GLM die Fehlerterme eine Verteilung aus der Klasse der exponentiellen Familie (z.B. Poisson, Gamma usw.) besitzen – für viele Fragestellungen im Gesundheitssystem das passendere Modell.

Vorteil beider Modelle ist, dass ein direkter funktionaler Zusammenhang zwischen den Eingabegrößen und der Zielgröße prognostiziert wird. Da die Eingabevariablen i.A. beeinflussbar sind, können aus dem (prognostizierten) kausalen Zusammenhang Optionen für die Steuerung der Zielgröße abgeleitet werden.

Neben den statistisch motivierten Verfahren gibt es die Klasse von Modellen ohne Verteilungsannahmen: die Verfahren des Maschine Learning, welche zunächst auf einer Trainingsmenge angelernt werden. Einer der einfachsten Vertreter dieser Klasse ist der **K-Nearest-Neighbour** Algorithmus. Dieser weist einem Punkt aus dem Eingaberaum den Durchschnitt der Zielgrößen, der k nächstgelegenen Punkte als Prognose zu. Auch die verschiedenen Formen von **Regressionsbäumen**, welche eine leichte Abwandelung der Entscheidungsbäume, Random Forests bzw. des Tree Boosting sind, gehören in diese Klasse, sowie die bereits erwähnten **Künstlichen Neuronalen Netze**.

Vorteil dieser Verfahren ist, dass oft komplexe, nichtlineare Zusammenhänge erkannt werden. Zudem müssen keine (potentiell falschen) Modellannahmen getroffen werden.

Allerdings benötigen diese Modelle häufig einen größeren Datensatz als die statistischen Verfahren, um gute Ergebnisse zu erzielen. Zudem haben die genannten Verfahren im Allgemeinen Probleme, sinnvolle Prognosen für Punkte zu liefern, die abseits der bekannten Trainingsmenge liegen. Darüber hinaus liefern diese Verfahren i.A. keinen direkten (erkennbaren) Zusammenhang zwischen Eingabe und Zielgröße.

Die folgende Abbildung gibt einen Überblick über einige bei der Modellauswahl für die metrische Prognose beachtenswerte Kriterien:

Metrische Prognose: **Die Methode ist geeignet für ...**	**Regressionsbaum**	**Random Forest (Regression)**	**Tree Boosting (Regression)**	**Neuronale Netze**	**lineare Regression**	**GLM**	**K-Nearest-Neighbour**
fehlende Merkmalswerte	+	+	+	-	-	-	-
nichtnumerische Merkmale	+	+	+	-	-	-	-
tendenziell knappe Trainingsdaten	+	+	+	-	+	+	-
nichtlineare Zusammnehänge	+	+	+	+	-	-	+
geringen Tunigbedarf	+	+	+	-	+	+	+
geringe Rechenkapazität	+	o	o	-	+	+	+
Das Prognosemodell ist ...							
anschaulich und interpretierbar	+	-	-	-	+	o	-
überschaubar (geringe Parameteranzahl)	+	-	-	-	+	+	-
stabil (z.B. Datenaktualisierung)	-	-	-	-	+	+	-
das mit der besten Vorhersage	-	o	+	+/-	-	-	-

d. Dimensionsreduktionsverfahren

Das **Dimensionsreduktionsverfahren** ist ein eher technisches Verfahren mit dem Ziel, multidimensionale Datensätze auf einige wenige Dimensionen zu reduzieren, ohne dabei viel an Informationsgehalt zu verlieren. Oftmals ist die Dimensionsreduktion weiteren Analyseschritten oder dem Lernen von Modellen wie z.B. Modellen zur Ereignisprognose (Kapitel 5a), Modellen zur metrischen Prognose (Kapitel 50) oder Empfehlungssystemen (Kapitel 5f) vorgelagert.

In der Krankenversicherungswirtschaft können Dimensionsreduktionsverfahren jetzt schon beispielsweise im Zuge von Bestandsanalysen, Leistungsdatenanalysen und Leistungsprognosen zum Einsatz kommen. Weiter an Bedeutung gewinnen werden

Dimensionsreduktionsverfahren, wenn die Daten- und Informationsdichte im Zuge der Digitalisierung weiter steigt.

Bei der **Hauptkomponentenanalyse (Principal Component Analysis – PCA)** werden beispielsweise die Hauptkomponenten als normalisierte Linearkombinationen der Merkmale im ursprünglichen Datensatz bestimmt. Die erste Hauptkomponente enthält den größten Anteil der Varianz. Jede weitere Hauptkomponente steht orthogonal zu den zuvor berechneten Hauptkomponenten. Die Hauptkomponenten sind unkorreliert.

Die Hauptkomponentenanalyse kann nur auf Datensätzen mit homogenen (numerischen) Merkmalen eingesetzt werden. **Generalised Low Rank Models (GLRM)** können allgemeiner mit heterogenen Merkmalen umgehen, also mit numerischen, booleschen, kategorischen, ordinalen und anderen Merkmalen.

e. Erfolgsmessung mittels Matching-Verfahren

Im Zusammenhang mit der Umsetzung von Maßnahmen oder Programmen wird oftmals auch eine Erfolgsmessung verlangt. In der GKV ist eine Erfolgsmessung z.B. für Wahltarife notwendig, um Einsparungen und Effizienzsteigerungen nachzuweisen. In der PKV könnte damit z.B. die Frage nach den Einsparungen bei der Durchführung von Gesundheitsmanagementmaßnahmen beantwortet werden.

Die Bildung einer Kontrollgruppe durch identische Zwillingspaare ist in hoher Dimensionalität ein unerreichbares Ideal, daher bedarf es alternativer Verfahren. Die Messung von beispielweise Einsparungen durch die nachträgliche Bildung einer geeigneten Kontrollgruppe kann in folgenden vier Schritten durchgeführt werden:

- **Schritt 1:** Auswahl eines Ähnlichkeitsmaßes
  Recht populär ist hier die Berechnung der Teilnahmewahrscheinlichkeit ("Propensity Score") mittels logistischer Regression. Dabei sollten alle zur Verfügung stehenden Kenntnisse wie Vertrags-, Kosten- und Morbiditätsinformation als erklärenden Variablen verwendet werden.
- **Schritt 2:** Auswahl einer Matching-Methode
  Beim Propensity-Score Matching (PSM) wird jedem Teilnehmer der Nicht-Teilnehmer mit dem ähnlichsten Propensity-Score zugeordnet. Zu diesem Standardverfahren sollten auch die zahlreichen Alternativen geprüft werden. Schritt 1. und 2. können auf einfache Weise mit Statistiksoftware umgesetzt werden, beispielweise mit der Funktion "matchit" in R.
- **Schritt 3:** Güte des Matching bewerten
  Durch die Berechnung des sogenannten "Standardized Bias" für jede Kovariate sowohl vor als auch nach dem Matching sowie die Verwendung eines entsprechenden "Jitter-Plots" für die Verteilung des Propensity Scores kann ein guter Eindruck über die Güte des Matching erfolgen.

- **Schritt 4:** Ergebnis analysieren
  Die in den Schritten 1. bis 3. gebildeten Kontrollgruppe wird nun zum Vergleich mit der Teilnehmergruppe herangezogen. Zur Beantwortung ökonomischen Fragestellungen sind Differenzschätzer, bei denen die Veränderung der entsprechenden Kennzahl der Teilnehmergruppe im Vergleich zur Kontrollgruppe gemessen wird, gut geeignet.

f. Empfehlungssysteme

Ein **Empfehlungssystem** ist ein System, welches individualisierte Empfehlungen erstellt oder dem Anwender in einer personalisierten Weise interessante oder nützliche Objekte aus vielen möglichen Optionen aufzeigt („a system that produces individualized recommendations as output or has the effect of guiding the user in a personalized way to interesting or useful objects in a large space of possible options" [Burke, 2002]). Ziel ist es dabei, auf der Grundlage von Daten und Informationen über eine Person („Benutzer") interessante oder passende Produkte bzw. Dienstleistungen („Objekte") zu identifizieren. Praktisch finden solche Empfehlungssysteme derzeit beispielsweise bei Amazon, Netflix und Spotify erfolgreich Anwendung.

In der Krankenversicherungswirtschaft vorstellbar sind Anwendungen insbesondere in der Vertriebsunterstützung oder im Zusammenhang mit Assistance-Leistungen (bspw. Arztempfehlungen). Aktuell besteht hier die Herausforderung, dass ein direkter Zugang zum Versicherungsnehmer in der für gezielte Auswertungen erforderlichen Form und damit eine entsprechend große Datengrundlage meist fehlt. Mit der Zunahme von Online-Vertriebswegen und Versicherungsnehmerportalen werden hier jedoch sukzessive neue Möglichkeiten entstehen.

In einem Empfehlungssystem werden typischerweise die folgenden drei Schritte unterschieden, die nacheinander durchlaufen werden:

- Benutzermodellierung,
- Generierung von Prognosen,
- Generierung von Empfehlungen.

Beim **inhaltsbasierten Filtering** werden Objekte empfohlen, die Objekten ähnlich sind, die der Benutzer bereits hoch bewertet hat. Die Ähnlichkeit zwischen Objekten wird feature-basiert bestimmt. Für die Modellbildung kommen je nach Zielsetzung Verfahren zur Ereignisprognose (Kapitel 5a) oder metrischen Prognose (Kapitel 50) zum Einsatz.

Beim **benutzerbasierten Collaborative Filtering** bestehen die Empfehlungen aus Objekten, an denen ähnliche Benutzer das größte Interesse haben. Zur Bestimmung der ähnlichen Benutzer werden die bisherigen Bewertungen der Benutzer herangezogen. Beim **objektbasierten Collaborate Filtering** hingegen werden Objekte empfoh-

len, die Objekten ähnlich sind, die der Benutzer bereits hoch bewertet hat. Die Ähnlichkeit zwischen Objekten wird hierbei ebenfalls auf Basis der bisherigen Bewertungen der Benutzer bestimmt. In beiden Varianten des Collaborative Filtering liegen also nur die bisherigen Bewertungen der Benutzer zu Grunde.

Beispielhaft sei generisch das Vorgehen des benutzerbasierten Collaborative Filtering erläutert. Zunächst werden beim benutzerbasierten Collaborative Filtering die Ähnlichkeiten zwischen dem aktuellen Benutzer und allen anderen Benutzern bestimmt (eine sehr einfache Form ist z.B. „ebenfalls 30-jährige weibliche Versicherte im Tarif XY"; relevante Vergleichskriterien können jedoch auch völlig andere, häufig auch viel weniger offensichtliche Kriterien sein). Dies kann online oder auch offline (d.h. vorab mit anschließender Speicherung der Ähnlichkeitsmatrix) geschehen. Im nächsten Schritt werden die Nearest Neighbours, also die ähnlichsten Benutzer, basierend auf den Benutzer-zu-Benutzer-Ähnlichkeiten bestimmt. Für die Prognoseberechnung im letzten Schritt wird die prognostizierte Bewertung als gewichteter Mittelwert der Bewertungen der Nearest Neighbours berechnet, wobei die Benutzer-zu-Benutzer-Ähnlichkeiten die Gewichte darstellen.

g. Modelltraining und Bewertung

Zur Vermeidung von **„Overfitting"** wird die Modellparameteroptimierung üblicherweise anhand von Teilstichproben durchgeführt. Dabei hat sich die 5- bis 10-fache Kreuzvalidierung bewährt. Das Modell wird mehrfach rotierend auf Basis des größeren Teils der Datensätze „trainiert" und anschließend mit den restlichen Datensätzen getestet.

Für die **Modellbewertung** sind bei einer Ereignisprognose die Fehlklassifikationsrate sowie der häufig in Prognosewettbewerben verwendete AUC-Wert („Area under ROC-Curve", vergleichbar mit Gini-Koeffizient) gut geeignet. Bei metrischen Prognosen wird gerne der mittlere absolute Fehler "MAE" oder auf Basis der Normalverteilungsannahme die mittlere quadratische Abweichung "RMSE" verwendet. Bei einem Teil der Software kann auch eine Nutzenmatrix verwendet und optimiert werden.

Ein interessantes Beispiel für die nötigen Schritte zur Erzielung einer guten binären Prognose ist die beim SAS Global Forum 2016 vorgestellte Studie von Akosa & Kelly "Application of Data Mining Techniques in Improving Breast Cancer Diagnosis"[6]. Dort ist auch der Nutzen einer zusätzlichen analytischen Datenaufbereitung (z.B. Hauptkomponententransformation) sowie von gestapelten Modellen gut erkennbar.

[6] siehe http://support.sas.com/resources/papers/proceedings16/9420-2016.pdf

6. Beispiel einer konkreten Modellierung zur Vorhersage von metrischen Leistungsausgaben in der Pflegeversicherung

Ziel dieses Abschnitts ist es auf Basis eines stark vereinfachten Beispiels die Vorgehensweise einer Datenanalyse modellhaft zu skizziert. Dazu soll für einen Tarifversicherten zum Termin t_0 die Leistungshöhe in Euro zu einem Zeitpunkt $t_0 + n$ vorhergesagt werden, d.h. es handelt sich um ein Problem aus der Klasse der metrischen Prognose.

Als konkretes Beispiel soll die Prognose der Leistungshöhe einer pflegebedürftigen Person dienen. Das Thema Pflege bietet sich aus verschiedenen Gründen zum Einstieg an, so ist der Pflegebegriff durch soziale Pflegeversicherung eindeutig definiert, es ist eine klare Abgrenzung durch Pflegestufen/Pflegegrade gegeben, die Bewertung in Abhängigkeit von Pflegestufe/Pflegegrad ist möglich, usw.

a. Datengrundlage

Für die Prognose wird von der Annahme ausgegangen, dass folgende Informationen zur Person und im Bestand des Versicherers vorliegen:

1) Unabänderliche Personendaten
   a) Geschlecht
   b) Geburtsdatum

2) Veränderliche Personendaten (Vektor mit Status der letzten n Jahre)
   a) PLZ des Wohnorts

3) Pflegedaten (Vektor mit Daten der letzten n Jahre)
   a) erste Pflegestufe/Pflegegrad
   b) Alter bei der ersten Pflegestufe/Pflegegrad

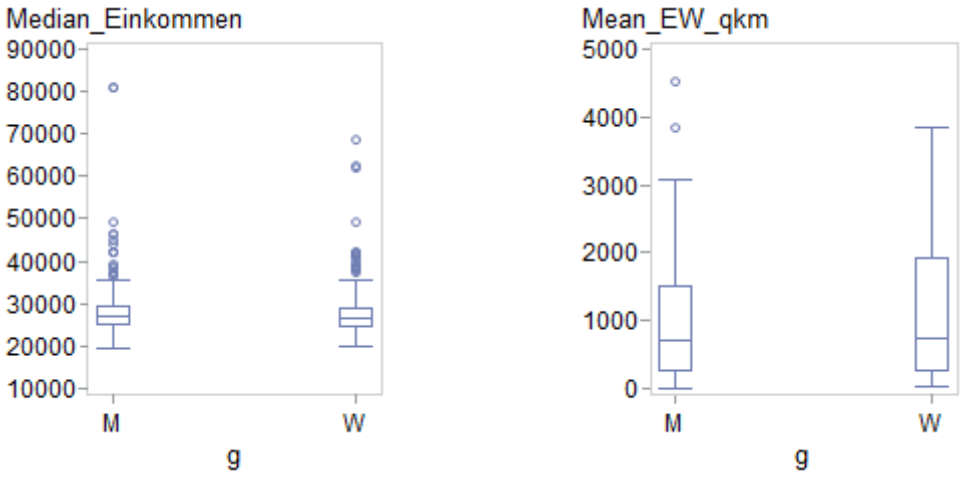
4) Leistungsdaten (Vektor mit Daten der letzten n Jahre)
   a) ambulante Leistungen
   b) stationäre Leistungen
   c) Zahnleistungen

Der Umfang der Daten ist bewusst sparsam gewählt, um das Modell möglichst einfach zu halten. Der verwendete Datensatz umfasst insgesamt 936 (verfälschte) Beobachtungen.

Die folgenden Auswertungen wurden mit den Softwareprodukten Enterprise Miner und Enterprise Guide von SAS erstellt. Neben dieser Software gibt es eine ganze Reihe weiterer geeigneter Softwareprodukte, kommerziell und opensource.

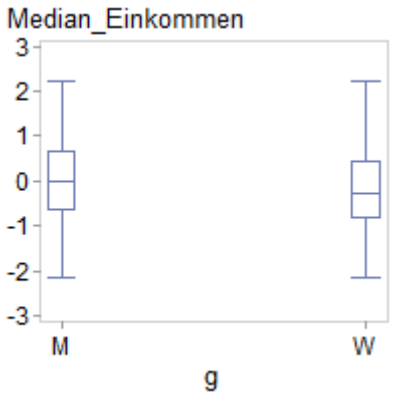
b. Datenvorbereitung

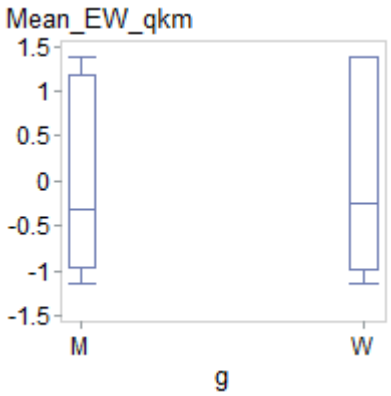
Im Folgenden wird die Datenaufbereitung anhand der Variablen PLZ skizzieren. Die PLZ als reiner Zahlenwert ist einer Region willkürlich zugeordnet worden und sollte über ihre Funktion als Regionenschlüssel keine weiteren Informationen enthalten. Um in der Datenanalyse regionale Merkmale zu berücksichtigen, wird mittels der PLZ der Datensatz um die geographischen Merkmale erweitert. Dabei sollen die Merkmale „ländliche Region/ Stadt“ und „reiche/arme Region“ an den Datensatz angefügt werden. Hierbei ist zu beachten, dass nicht alle Modell kategorische Variablen verarbeiten können. Alternativ zu kategorischen Merkmalen können oftmals stetige Merkmale wie „durchschnittliche Bevölkerungszahl pro qkm“ oder das „Median-Einkommen pro Haushalt“ verwendet werden. Allerdings sind stetig verteilte Größen weniger Robust gegen Ausreißer, als kategorische Variablen. Auch die Verteilung der beiden Merkmale im vorliegenden Datensatz besitzen Ausreißer, wie nachfolgenden Grafiken zeigen:



Für die Datenanalysen könnten vor allem die extremen Ausreißer zu Problemen führen. Daher wird der Datensatz um diese bereinigt. Der einfachste Ansatz dafür ist es beim Einkommen alle Werte, die außerhalb eines festzulegenden Intervalls, z.B. [20.000, 35.000] Euro liegen, auf die obere bzw. untere Grenze Intervallgrenze zu setzen. Analog werden die Werte der durchschnittlichen Einwohnerzahl auf das Intervall [100, 1500] „gestutzt“.

Die beiden Variablen besitzen nun eine um den Faktor 10 verschiedene Größenordnung. Solche unterschiedlichen Skalierungen können unter Umständen zu Problemen zu Modellinstabilitäten führen. Um solche Effekte zu vermeiden, werden die beiden Variablen auf den Erwartungswert Null und die Standardabweichung 1 normiert. Das Ergebnis sieht wie folgt aus:

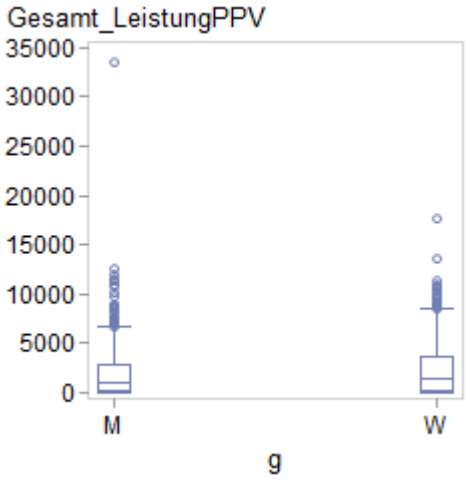




Weitere Datenaufbereitungsschritte schließen sich an, auf die aber nicht weiter eingegangen werden soll. Als Ergebnis der Datenaufbereitung und Merkmalsgenerierung ergeben sich die folgenden Variablen:

Prediktorvariable	**Beschreibung (Summen pro Person)**
Alter_Erst_Pflege	Alter bei der ersten auftretenden Pflegestufe
Erste_Pflegestufe	Erste Pflegestufe
KV_Leistung_IntA1	Summe ambulanter Leistungen ein Jahr vor der Pflege
KV_Leistung_IntA2	Summe ambulanter Leistungen zwei Jahre vor der Pflege
KV_Leistung_IntA3	Summe ambulanter Leistungen drei Jahre vor der Pflege
KV_Leistung_IntA_all	Summe aller ambulanten Leistungen
KV_Leistung_IntS1	Summe stationärer Leistungen ein Jahr vor der Pflege
KV_Leistung_IntS2	Summe stationärer Leistungen zwei Jahre vor der Pflege
KV_Leistung_IntS3	Summe stationärer Leistungen drei Jahre vor der Pflege
KV_Leistung_IntS_all	Summe aller stationären Leistungen
KV_Leistung_IntZ1	Summe Zahn-Leistungen ein Jahr vor der Pflege
KV_Leistung_IntZ2	Summe Zahn-Leistungen zwei Jahre vor der Pflege
KV_Leistung_IntZ3	Summe Zahn-Leistungen drei Jahre vor der Pflege
KV_Leistung_IntZ_all	Summe aller Zahn-Leistungen
Mean_EW_qkm	Mittlere Einwohnerzahl pro Quadratkilometer unter dieser PLZ (falls vorhanden)
Median_Einkommen	Median des Einkommens unter dieser PLZ (falls vorhanden)
g	Geschlecht

Die zu schätzende Zielgröße sind die Pflegeleistungen in den ersten 6 Jahren nach Zuordnung der ersten Pflegestufe. Im Beispieldatensatz ergibt sich folgende Verteilung der Zielgröße für Frauen und Männer:

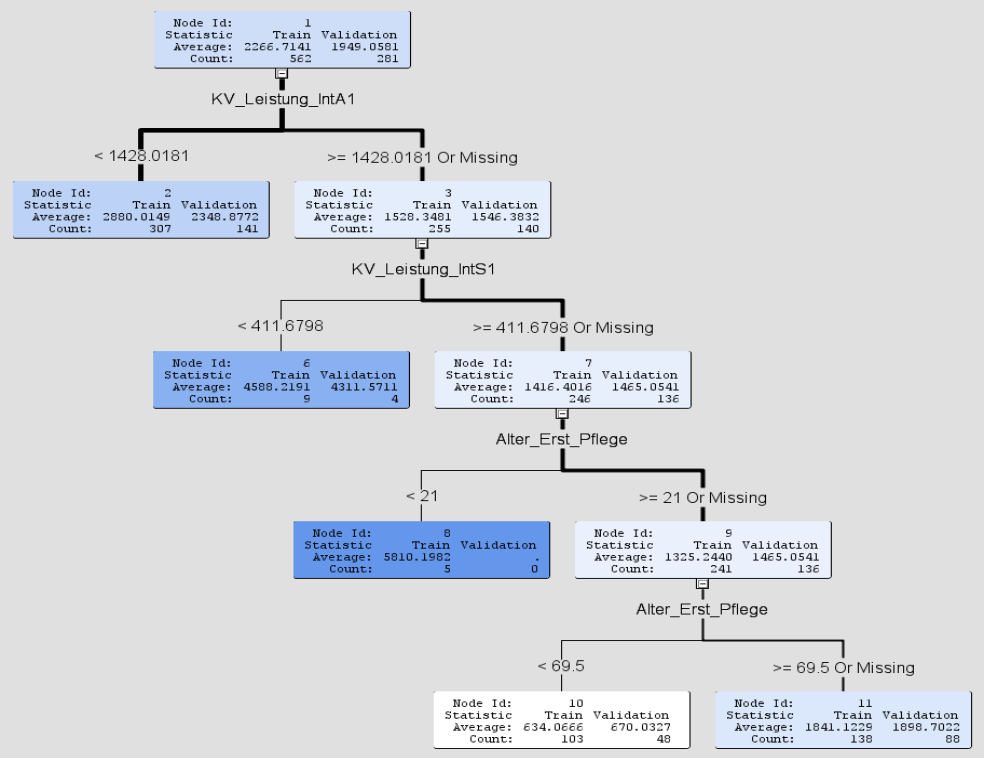


Es zeigt sich, dass der Median und die Quantile bei den Frauen größere Werte annehmen, als bei den Männern. Der eine extreme Ausreißer bei den Männern und die beiden Ausreißer bei den Frauen könnten in der Analyse Probleme mit sich bringen und sollten aus dem Datensatz ausgeschlossen werden.

c. Modelltraining und -bewertung

Zum Training und zur Bewertung verschiedener Modelle wird der Beispieldatensatz in 3 Teildatensätze zerlegt, den Trainings- (60% der Datenpunkte), Validation- (30%) und Testdatensatz (10%). Der Trainingsdatensatz dient zum Training der einzelnen Modelle mit unterschiedlichen Parametereinstellungen. Im Anschluss daran die Parametreinstellung ermittelt, welche die beste Approximation der Zielgröße auf dem Validation-Datensatz liefert. Danach werden die verschiedenen Modelle (mit den besten Parametereinstellungen) miteinander verglichen. Dafür kann nicht den Wert auf dem Validationsdatensatz verwendet werden, da dieser durch die Suche nach der optimalen Parametereinstellung schon für das Training der Modelle verwendet wurde. Durch die intensive Suche nach der besten Parametereinstellung könnte sich das „Siegermodell“ zu stark auf den Validationsdatensatz angepasst haben. Auf anderen Datensätzen könnte dieses „Modell“ dann ungeeignet sein. Um dies zu erkennen braucht es den dritten, unabhängigen Testdatensatz. (In der Literatur gibt es eine Vielzahl von Artikeln, die sich mit der Problematik der Modellbewertung und Validierung beschäftigen und es empfiehlt sich, sich einen Überblick über diese Thematik zu verschaffen.)

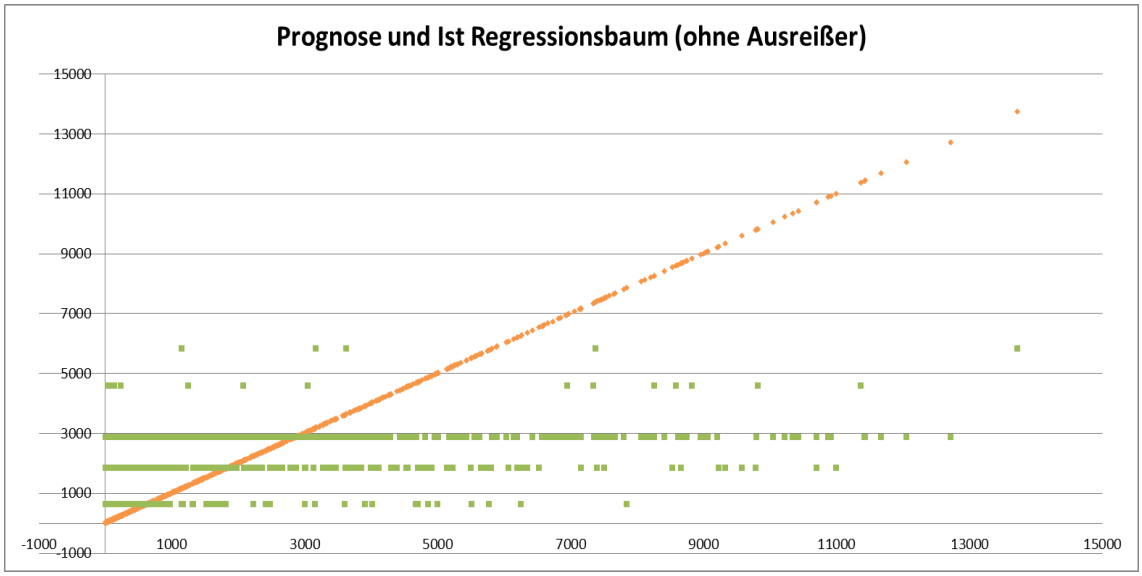
Als Beispielmodell wird die Klasse der Regressionsbäume gewählt, welche die Zielfunktion mittels einer Art Treppenfunktion approximieren. Dazu zerlegt der Algorithmus den Raum der Prediktorvariablen mit Hilfe von Splitting-Rules in Teilräume. Auf diesen Teilräumen wird die Zielfunktion durch einen geeigneten Schätzer (zumeist der Durchschnitt der Zielfunktion über die Datenpunkte des Teilraums) approximiert. Die nächste Grafik zeigt einen auf dem Datensatz trainierten Regressionsbaum:



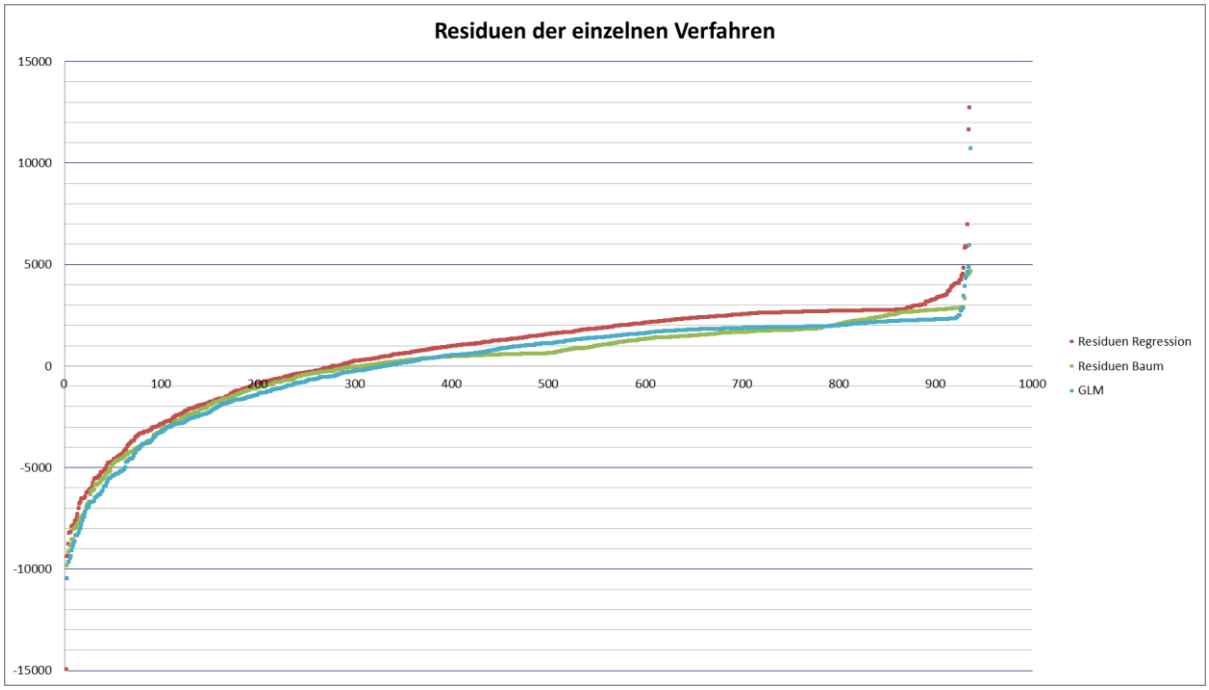
Der Baum besitzt insgesamt 5 Blätter, und somit 5 verschiedene Vorhersagewerte. Zwei Blätter (6 und 8) sind jedoch sehr dünn besetzt, so dass die relevante Prognose nur durch die Blätter 2, 10 und 11 erfolgt. Als Prognosewert wird der Durchschnitt aller Trainingsvariablen im jeweiligen Blatt verwendet, d.h. z.B. für Blatt 10 wird die Zielfunktion mit 634,07 € geschätzt.

Neben der Prognose der Zielfunktion liefert der Baum auch eine Vorhersage für die Relevanz der einzelne Prediktorvariablen. Als wesentlichste erklärende Variable hat der Baum die „ambulanten Leistungen im letzten Jahr vor der Pflegestufe“ ausgemacht, gefolgt von den „stationären Leistungen im letzten Jahr vor der Pflegestufe“ und dem „Alter bei der ersten Pflegestufe“. Die übrigen Variablen haben nach dem Baum weniger Relevanz.

Um ein Gefühl für die Güte des Modells zu erhalten, werden die realen Zielgrößen über sich selbst (orange „Punktlinie“) und den vom Modell prognostizierten Wert geplottet. Ein perfektes Modell würde die orange Punktlinie überdecken. Wie erwartet liefert das sehr einfache Modell keine gute Approximation. Im Plot sind die fünf möglichen Prognosewerte des Regressionsbaums zu erkennen, deren Anzahl für eine gute Näherung viel zu gering ist.



Nachdem der „Regressionsbaum“ trainiert wurde, soll dieses Modell mit anderen verglichen werden. Als Vergleichsmodelle werden ein lineares Regressionsmodell und ein GLM trainiert. Die Details des Trainings und der Parameterfindung sollen an dieser Stelle nicht vertieft werden. Zur Beurteilung der Güte der verschiedenen Modelle untereinander wird der nicht (zum Training/ zur Parameterfindung) verwendete Testdatensatz verwendet. Auf diesen Datensatz ist ein Maß für die Qualität der einzelnen Modelle zu wählen. Im Folgenden wird dafür die „Wurzel des durchschnittlichen quadratischen Fehlers“ der Modelle auf dem Testdatensatz verwendet. Der Testdatensatz beinhaltet knapp 100 Datenpunkte. Auf diesen Datenpunkten beträgt die Wurzel des durchschnittlichen quadratischen Fehlers 2.919 für die lineare Regression. Im Vergleich dazu ist der Fehler des GLM mit 2.741 niedriger. Der Wert des Regressionsbaums liegt mit 2.681 noch etwas darunter. Daher wäre der Regressionsbaum den beiden anderen Modellen vorzuziehen. Bei der Betrachtung der (nach der Größe geordneten) Residuen der einzelnen Modelle werden die (leichten) Vorzüge des Regressionsbaums gegenüber den anderen Modellen ersichtlich.



Dennoch ist festzustellen, dass keines der Modelle befriedigende Ergebnisse geliefert hat. Die schlechte Performance könnte darauf hindeuten, dass eventuell wesentliche erklärende Variablen (wie z.B. Diagnosen) in den Modellen nicht verwendet wurden. Auch könnten die Modelle an sich ungeeignet sein, um den (potentiellen) Zusammenhang zu erklären, z.B. könnte das Problem hochgradig nicht linear sein, was die lineare Regression an sich als Modell ausschließen würde. Das Problem beim Regressionsbaum ist dagegen die geringe Datenbasis, da dieses Verfahren im Allgemeinen erst bei wesentlich größeren Datensätzen gut funktioniert.

7. Literaturhinweise

1. Methodik maschinelles/statistisches Lernen:

James G et al., An Introduction to Statistical Learning with Applications in R, Springer 2013
https://lagunita.stanford.edu/courses/HumanitiesSciences/StatLearning/Winter2016/about

Hastie T, Tibshirani R, Friedman J, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition, Springer 2009
http://statweb.stanford.edu/~tibs/ElemStatLearn/

Goodfellow I, Bengio Y, Courville A, Deep Learning (Adaptive Computation and Machine Learning), MIT Press 2017
http://www.deeplearningbook.org/

2. Matching-Verfahren:

Kuss O, Blettner M, Börgermann J, 2016, Propensity Score - eine alternative Methode zur Analyse von Therapieeffekten. Deutsches Ärzteblatt 5. Sept. 2016.
https://www.aerzteblatt.de/archiv/181706/Propensity-Score-eine-alternative-Methode-zur-Analyse-von-Therapieeffekten

Freytag A et al., 2016, Effekte hausarztzentrierter Versorgung, Deutsches Ärzteblatt 25. November 2016
https://www.aerzteblatt.de/archiv/183908/Effekte-hausarztzentrierter-Versorgung

Stuart EA, 2010, Matching methods for causal inference: A review and a look forward. Stat. Sci 25(1):1-21
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2943670/pdf/nihms200640.pdf

Gensler S, Skiera B, Böhm M, 2005, Einsatzmöglichkeiten der Matching Methode zur Berücksichtigung von Selbstselektion, JfB 55: 37–62.
https://www.researchgate.net/publication/226093740_Einsatzmglichkeiten_der_Matching_Methode_zur_Bercksichtigung_von_Selbstselektion

3. Weiteres Material:

Friedrich Loser, Okt. 2016, Data Science/-Mining/-Machine Learning mit R für Aktuare: Eine kurze, praktische Einführung
https://aktuar.de/interner-bereich/vereinsinterna/davvorort/2016-12-07_Vortrag_Loser.pdf

8. Anhang: Anmerkungen zum Datenschutz in Bezug auf die Nutzung von personenbezogenen Daten, insbesondere von Gesundheitsdaten und Anwendung der Analyseergebnisse

Personenbezogene Daten sind Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren Person, Gesundheitsdaten sind eine besondere Art von personenbezogenen Daten (§ 3 Abs. 1 und 9 BDSG), für die Sonderregelungen nach §§ 28 Abs. 6-9, 29 Abs. 5 BDSG gelten.

Das Erheben, Verarbeiten und Nutzen von personenbezogenen Daten und vor allem von Gesundheitsdaten durch ein privates Unternehmen für eigene Geschäftszwecke ist nur dann zulässig, wenn der Betroffene einwilligt oder ein gesetzlich definierter Ausnahmefall vorliegt, § 4 Abs. 1 BDSG.

Neben diesem Verbot mit Erlaubnisvorbehalt sind weitere Grundprinzipien im Datenschutz zu beachten: Direkterhebung, Datensparsamkeit, Datenvermeidung, Zweckbindung, Transparenz und Erforderlichkeit. Zum Konzept von Big Data, bei dem möglichst viele Daten für alle denkbaren Zwecke zu verwenden, wird hierzu ein Widerspruch gesehen. [7]

In der Regel liegt eine Einwilligung bei Vertragsschluss vor über die Erhebung, Speicherung und Nutzung der mitgeteilten Gesundheitsdaten soweit dies zur Durchführung oder Beendigung des Versicherungsvertrags erforderlich ist. Während der Vertragslaufzeit können weitere Einwilligungen eingeholt werden.

Folgende Sachverhalte sind genau zu unterscheiden bei der Frage, ob die Verwendung der Daten von der Einwilligung umfasst sind.

- Eingereichte Belege (Arztrechnungen, Rezepte für Arzneimittel, Krankenhausrechnungen)
    - Selektion der einzelnen Belege ohne Verknüpfungen, z. B. nach einem ICD-10-Code: Der Versicherer muss sich bei einer Einwilligung (wie oben beschrieben) die Frage stellen, was unter „Durchführung des Versicherungsschutzes“ verstanden wird. Dies wird in der Regel durch den Versicherungsvertrag (AVB, Tarife) beantwortet.
    - Verknüpfung verschiedener eingereichter Belege, z. B. Arztrechnungen mit verschiedenen ICD-10-Codes und Arzneimittelrezepte mit bestimmten Pharmazentralnummern: Eine genaue Prüfung ist erforderlich, ob

[7] Becker/Schwab: Big Data im Gesundheitswesen – Datenschutzrechtliche Zulässigkeit und Lösungsansätze, ZD 2015, 151 (153)

dies noch unter den Vertragszweck der Kostenerstattung für den Versicherungsfall fällt, wenn nicht die Belege für einen Versicherungsfall betrachtet werden, sondern diese verschiedenen Versicherungsfälle betreffen, und verknüpft werden.

- Verknüpfung mit anderen Daten aus der Krankenversicherung, z. B. Daten aus der Antragsprüfung wie BMI, Raucher/Nichtraucher, Beruf, Wohnort. Diese Daten hat der Versicherte in der Regel zum Zweck der Antragsprüfung, der richtigen Tarifierung oder für die Zustellung von Post mitgeteilt, jedoch nicht zum Zweck der Kostenerstattung. Daher ist bei einer Verknüpfung mit Belegen, die der Versicherte zu Kostenerstattung einreicht, der Zweck der Einwilligung zu prüfen.
- Verknüpfung mit Daten aus anderen Sparten (LV, BU). Es ist zu prüfen, auf welche Daten sich die Einwilligung bezieht. Andere Sparten können die personenbezogenen Daten, die sie gespeichert haben, auch nur aufgrund einer Einwilligung zur Verfügung stellen.
- Verknüpfung mit Daten aus dem Internet, z. B. Gesundheitsforen oder Facebook.

Wenn keine Einwilligung der Betroffenen über die Verwendung dieser Daten vorliegt, könnte der Erlaubnistatbestand nach § 28 Abs. 6 Nr. 2 BDSG vorliegen, wenn die Auswertung für eigene geschäftliche Zwecke genutzt werden soll. Dann muss es sich um Daten handeln, die der Betroffene offenkundig öffentlich gemacht hat. Es muss feststehen, dass die Veröffentlichung durch den Betroffenen tatsächlich akzeptiert wurde. Dies kann bei Daten, die Nutzer auf Plattformen abgeben, die eine gewisse vorherige Anmeldung oder Authentifizierung erfordern, nicht ohne weiteres angenommen werden.[8]

Wenn keine Einwilligung vorliegt, könnten verschiedene gesetzliche Erlaubnistatbestände zur Anwendung kommen. Während im Sozialgesetzbuch V speziell für die Krankenkassen Erlaubnistatbestände geregelt sind (z. B. § 197 a SGB V oder § 137 f SGB V), gibt es für die private Krankenversicherung keine entsprechenden Vorschriften im Versicherungsvertragsgesetz. Daher muss auf das BDSG zurückgegriffen werden sowie – falls ein Unternehmen beigetreten ist – auf den Code of Conduct. Es ist für jeden Anwendungsfall gesondert zu prüfen, ob ein Erlaubnistatbestand und welcher Erlaubnistatbestand in Frage kommt. Geht es zum Beispiel um die Aufdeckung von

[8] Becker/Schwab: Big Data im Gesundheitswesen – Datenschutzrechtliche Zulässigkeit und Lösungsansätze, ZD 2015, 151 (153f.)

Missbrauch von Versicherungsnehmern, kann der Erlaubnistatbestand nach § 28 Abs. 1 Nr. 1 BDSG vorliegen. Dabei ist darauf zu achten, welche Daten von der Big-Data Anwendung erfasst sind, weil sich die Regelungen für Gesundheitsdaten in § 28 Abs. 6 BDSG befindet. Bei Disease Management Programmen könnte der Erlaubnistatbestand des § 28 Abs. 7 BDSG herangezogen werden.

Lösungsansätze

Welche Lösungsansätze darüber hinaus zulässig sind, muss nach dem Anwendungsfall bestimmt werden.

- Anonymisierung: Die Anwendbarkeit des Datenschutzrechts wird nach § 3 Abs. 6 BDSG dann ausgeschlossen, wenn personenbezogene Daten in der Art verändert werden, dass die Einzelangaben über persönliche Verhältnisse nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können[9].
- Pseudonymisierung: Grundsätzlich ist die Anonymisierung vorzuziehen.
- Einwilligung: Bei einer Einwilligung auch in Big Data-Analysen ist die Herausforderung, die mit der Datenverarbeitung verfolgten Zwecke verständlich und so umfassend wie möglich darzustellen[10].

Ausblick auf die Rechtslage ab 25. Mai 2018

Die EU-Datenschutzgrundverordnung tritt am 25. Mai 2018 in Kraft. Der Entwurf eines neuen BDSG[11] liegt vor. Auch nach diesen Vorschriften gilt der Grundsatz des Verbots mit Erlaubnisvorbehalt.

Nach der EU-GVO wird eine Weiterverarbeitung von personenbezogenen Daten ohne erneute Einwilligung ausdrücklich ermöglicht, wobei hierfür enge Grenze gelten, Art. 5 Abs. 1 Buchst. b i.V.m. Art. 6 Abs. 4 DS-GVO. Ob diese Erlaubnisnorm auch für Gesundheitsdaten gilt, ist noch nicht abschließend geklärt (der GDV bejaht dies in „Unverbindliche Orientierungshilfe für die Praxis im Versicherungsunternehmen“ GDV-

[9] Ohrtmann/Schwiering, Big Data und Datenschutz – Rechtliche Herausforderungen und Lösungsansätze, NJW 2014, 2984

[10] Ohrtmann/Schwiering, Big Data und Datenschutz – Rechtliche Herausforderungen und Lösungsansätze, NJW 2014, 2984

[11] Gesetzentwurf in der Fassung des Bundestagsbeschlusses vom 24.4.2017 (BT Drs. 18/11325 und 18/12084).

Rundschreiben vom 30.01.2017). Allerdings wird auch vertreten, dass Gesundheitsdaten von der Erlaubnisnorm nicht umfasst sind. Dabei wird u. a. auf § 24 Abs. 2 BDSG-Entwurf verwiesen, der nur die Weiterverarbeitung zulässt, wenn sie zur Abwehr von Gefahren für die staatliche oder öffentliche Sicherheit oder zur Verfolgung von Straftaten erforderlich ist oder sie zur Geltendmachung, Ausübung oder Verteidigung rechtlicher Ansprüche erforderlich ist.