DAV

DEUTSCHE
AKTUARVEREINIGUNG e.V.

Ergebnisbericht des Ausschusses Lebensversicherung

Sterblichkeitsanalyse des NHANES-Datensatz mit Data Science-Methoden

Köln, 21. September 2022

Präambel

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

Zusammenfassung

Der Ergebnisbericht behandelt Fragestellungen und Beispieldaten zu Methoden aus dem Bereich Data Science und betrifft Aktuare in der Rolle als Verantwortlicher Aktuar, Sachverständiger, Aktuar bei einer Wirtschaftsprüfungsgesellschaft, Versicherungsmathematischen Funktion etc. bei der Ausführung aktuarieller Aufgaben insbesondere im Rahmen der Produktentwicklung und des aktuariellen Controllings.

Zur Veranschaulichung von Verfahren aus der Data Science wird hierzu ein Datensatz auf der Grundlage öffentlich verfügbarer und nutzbarer Daten aus den USA hergeleitet, an dem typische Fragestellungen zur Analyse biometrischer Risiken von Beständen beispielhaft durchgeführt werden können. Eine gute und systematische Aufbereitung von Daten ist der erste Schritt in Data-Science-Analysen. Die in diesem Bericht hergeleiteten Daten lassen sich vielfältig auswerten und geben ein komplexes Bild der untersuchten Personen wieder. Die hergeleiteten Daten werden in einem Folgebericht näher analysiert, können aber auch von allen Aktuaren als Grundlage für eigene Analysen genutzt werden.

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 Lebensversicherung am 21. September 2022 verabschiedet worden.

---

[1] Der Ausschuss dankt der Arbeitsgruppe Big Data in der Lebensversicherung ausdrücklich für die geleistete Arbeit, namentlich Stefan Heyers, Dominique Achard, Dr. Fabian Bohnert, Johanna Cheng, Andreas Döring, Thomas Gehling, Henri Grunder, Dr. Michael Hoffmann, Sven Rehmann, Hanna Speller, Katja Vogt, Dr. Frank Wittemann.

[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 durch eine(n) qualifizierte(n) Aktuar DAV/Aktuarin DAV getroffen werden.

Inhaltsverzeichnis

1. Einleitung

Biometrische Datenanalysen gehören zu den wichtigsten Anwendungen von Data-Science-Methoden in der Lebensversicherung. Durch die Digitalisierung und Deregulierung wird die Tarifierung biometrischer Produkte von mehr Dimensionen abhängig. Die klassischen univariaten Methoden reichen hier nicht mehr zur Analyse aus. Aus diesem Grund betrachtet der Bericht erweiterte Analysemöglichkeiten am Beispiel von Sterblichkeitsuntersuchungen.

Die Verwendung deutscher Versichertendaten gestaltet sich aus Gründen des Datenschutz- und Kartellrechts schwierig. Daher führen wir die Sterblichkeitsanalysen auf dem öffentlichen NHANES-Datensatz[3] durch, einem seit langem bekannten zur Verfügung gestellten Datensatz aus den USA von hoher Qualität. Der Hintergrund und Aufbau des Datensatzes werden in Kapitel 3 detailliert beschrieben.

Nach der Datenakquise stellen wir in Kapitel 4 kurz unsere Zusammenarbeit und die dabei verwendeten Tools vor. Danach folgen wir weiter den typischen Schritten des Data-Science-Workflows: der Datenaufbereitung und der Exploration. Während Kapitel 5 eine Beschreibung der Skripte bietet sowie auf deren Systemvoraussetzungen eingeht, werden in Kapitel 6 die zur Datenexploration angewendeten Methoden und deren Ergebnisse aufgezeigt.

In Kapitel 7 zeigen wir erste Ansätze des letzten Data-Science-Workflow-Schritts, der Analyse der Daten mit unterschiedlichen Verfahren. Genauer gehen wir hier auf die klassische Analyse sowie Regressionsverfahren ein. Dieser Abschnitt zeigt exemplarische erste Ergebnisse der eigentlichen Data-Science-Analysen. Ausführliche Ergebnisse und die zugehörigen Methodendarstellung inklusive der Skripte wird die Arbeitsgruppe in einem weiteren Bericht veröffentlichen.

Die Fokussierung dieses Berichtes auf eine Erzeugung von auswertbaren NHANES-Daten mag auf den ersten Blick überraschen. Erstaunlicherweise scheint es bisher keine vergleichbaren Quellen von Skripten oder Daten zu geben, die es ermöglichen würden, direkt Sterblichkeitsanalysen auf den NHANES-Daten durchzuführen. Deshalb ist die Arbeitsgruppe der Überzeugung, dass die Bereitstellung der Skripte zur Herleitung eines umfangreichen Datensatzes einen hohen Wert an sich hat, um die vielfältigen Variablen des NHANES-Datensatzes (vergleiche Abschnitt 3.2) zu untersuchen. Gleichzeitig wird mit diesem Datensatz geeignetes Testmaterial zur Verfügung gestellt, die verschiedenen Data-Science-Methoden zu testen. Von einer direkten Veröffentlichung der aufbereiteten NHANES-Daten hat die Arbeitsgruppe aus Urheberrechtsgründen abgesehen.

Die Ergebnisse zur Sterblichkeit aus dem NHANES-Datensatz sind nicht direkt auf andere Länder wie Deutschland übertragbar. Dennoch sind z. B. die ersten ge-

[3] NHANES-Datensatz: langjährige Datenerhebung der CDC (Center for Disease Control and Prevention, US) in den USA, https://www.cdc.gov/nchs/nhanes/index.htm, vgl. Abschnitt 3

zeigten Ergebnisse qualitativ und auch quantitativ interessant, und können geeignet angepasst auch als Quelle für andere Länder herangezogen werden.

Neben dem Datensatz, den Skripten zur Herleitung und folgend auch exemplarischen Ergebnissen zur Sterblichkeit geben auch die verwendeten Tools einen ersten Einblick in die neuere, professionellere Methodik einer Datenanalyse heutzutage: Dies gilt insbesondere für kleine Gruppen von bis zu 10 Personen.

Die Arbeitsgruppe hofft auf interessierte Kenntnisnahme und steht für Rückfragen jederzeit gerne zur Verfügung.

2. Umfang des Berichtes und Begriffsdefinition

Der NHANES-Datensatz ähnelt bezüglich Sterblichkeitsuntersuchungen stark den vorhandenen Daten einer Lebensversicherung: Beim Underwriting (hier: Studienbeginn) werden Daten zur Person erfasst. Diese Daten variieren in Umfang und Qualität mit dem Jahr des Studienbeginns. In den Folgejahren sind außer Todesfälle keine biometrischen Veränderungen der Personen bekannt.

- Aufgrund dieser Ähnlichkeiten und des zu erwartenden Nutzens des Datensatzes für Actuarial Data Science Projekte legt dieser Bericht und der veröffentlichte Quellcode besonderen Wert auf eine strukturierte, transparente und weiterentwickelbare Datenherleitung.
- Der Bericht und der Folgebericht beleuchtet Einsatzmöglichkeiten von Methoden aus den Bereichen Big Data und Künstlicher Intelligenz (insbesondere Machine Learning).
- Es werden vorwiegend Einsatzmöglichkeiten dargestellt, die charakteristisch für das Versicherungsgeschäft sind. Auf Prozesse, die bereits in anderen Branchen etabliert sind, wird nicht näher eingegangen - z. B. Analyse von Cross-Selling Möglichkeiten oder die Zuweisung der Post an zuständige Mitarbeiter im Unternehmen, vgl. aber den Bericht der DAV-Arbeitsgruppe Big Data [1]

2.1. Begriffsdefinitionen

Im Folgenden werden einige wichtige Begriffe im Umfeld von Künstlicher Intelligenz erläutert, die dem Ergebnisbericht „Anwendung von künstlicher Intelligenz in der Versicherungswirtschaft" [9] entnommen sind.

- **Data Science:** Unter Data Science verstehen wir die Wissenschaft, die sich mit Daten im Allgemeinen beschäftigt. Das umfasst den gesamten Prozess im Umgang mit Daten. Beginnend mit der Erhebung und Erfassung von Daten, der Aufbereitung und Speicherung, der Verarbeitung und Auswertung bis hin zum Data Mining. Bei der Erhebung und Verarbeitung der Daten sind technische Fragen des Datenmanagements ebenso relevant wie die Prinzipien der Datenschutzbestimmungen.
- **Data Analytics / Data Mining:** Unter Data Analytics verstehen wir die Gewinnung von (neuen) Informationen aus Daten unter einer spezifischen Fragestellung. Dies wird häufig auch als Data Mining bezeichnet. Wir bezeichnen mit Data Mining jedoch den gesamten Prozess, wie er in den einschlägigen Data Mining Prozessmodellen definiert ist. Data Mining umfasst somit insbesondere die Phasen: Datenbeschaffung, Datenaufbereitung, Modellierung, Data Analytics, Modellvalidierung, Evaluierung.
- **Actuarial Data Science:** Da die spezifischen Fragestellungen für uns stets aktuarieller Natur sind, oder allgemeiner das Prinzip Versicherung betreffen, sprechen wir von Actuarial Data Science (kurz ADS). Actuarial Data Science beschäftigt sich also mit der Erhebung, Erfassung, Verarbei-

tung und Auswertung versicherungsspezifischer Daten unter einer aktuariellen Fragestellung.

- **Machine learning (ML):** Die Auswertung bzw. Analyse der Daten (Data Analytics) erfolgt meist mit Methoden des maschinellen Lernens. ML steht im Gegensatz zur klassischen Programmierung, bei der ein (vorab bekannter und programmierter) Algorithmus für eine Dateneingabe abgearbeitet wird. Beim ML ist vorab kein Algorithmus bekannt (und meist auch nicht danach). Basierend auf einem Lernverfahren wird stattdessen anhand der Eingangsdaten ein Modell trainiert. Man sagt auch das Modell lernt aus den Daten. Wenn der Lernvorgang erfolgreich war, lässt sich das trainierte Modell auf neue Daten anwenden (überwachtes Lernen) oder das Modell liefert intrinsische Muster in den Daten und damit neue Informationen über die Daten (unüberwachtes Lernen).
- **Statistical learning (SL):** Während beim überwachten maschinellen Lernen die zentrale Aufgabe in der Vorhersage liegt, steht beim statistischen Lernen die Interpretation im Fokus. Darüber hinaus ist die Berücksichtigung von Zufallsfehlern und die Quantifizierung von Unsicherheit zentrales Instrument bei statistischen Lernverfahren. Unter Unsicherheit werden dabei sowohl Modellunischerheit als auch "einfache" Zufallsschwankungen verstanden. Eine scharfe Abgrenzung der Gebiete ist nicht möglich, da der Übergang zwischen den Bereichen nahezu fließend ist. Oftmals spricht man daher allgemein von Maschinellem Lernen, auch wenn man statistische Verfahren benutzt.
- **Künstliche Intelligenz:** Der Begriff künstliche Intelligenz bezeichnet den Versuch, menschenähnliche Entscheidungsstrukturen maschinell (in der Regel mit einer Software) nachzubilden. Konkret geht es darum, eine Maschine zu bauen, die eigenständig Aufgaben bearbeiten oder Probleme lösen kann. Machine Learning ist ein Teilbereich der Künstlichen Intelligenz.

Eine Einführung in die Themen, die in diesem Bericht behandelt werden, wird u. a. im Bericht der DAV „Big Data in der Lebensversicherung“ [1] und dem Ergebnisbericht „Anwendung von künstlicher Intelligenz in der Versicherungswirtschaft“ [9] sowie den dort genannten Literaturverweisen gegeben. Die Lektüre des vorliegenden Berichts setzt ein Grundverständnis der dort dargestellten Begriffe voraus.

Thematisch ist der vorliegende Bericht demnach am ehesten dem Bereich des Machine Learnings zuzuordnen. Der Datensatz mit 65.000 beobachteten Personen gehört sicherlich nicht zu „Big Data“. Neuronale Netze zur Sterblichkeitsanalyse hat die Arbeitsgruppe nicht näher betrachtet. Hierzu gibt es u. a. den thematisch ähnlichen Anwendungsfall der DAV-Arbeitsgruppe Actuarial Data Science „Neuronale Netze treffen auf Mortalitätsprognose“ [8].

3. Das NHANES Studienprogramm

3.1. Hintergrund

"National Health and Nutrition Survey" (NHANES) ist ein Studienprogramm des CDC (Centers for Disease Control and Prevention) zur Beurteilung des Gesundheits- und Ernährungszustandes von Kindern und Erwachsenen in den Vereinigten Staaten. Nur die Daten von Erwachsenen ab dem 18. Lebensjahr werden veröffentlicht.

Unter die hier betrachteten Daten fallen zwei leicht unterschiedliche Programme in den Bereich von NHANES: NHANES III und NHANES Continuous, wobei das erste von 1988 bis 1994 und das zweite von 1999 bis heute durchgeführt wurde.

Unabhängig davon, ob es sich um NHANES III oder NHANES Continuous handelt, ist der Ablauf wie im folgenden Absatz erläutert.

Alle drei Jahre (NHANES III) bzw. zwei Jahre (NHANES Continuous) wird eine neue Kohorte untersucht: Für jedes Mitglied einer gegebenen Kohorte werden Daten aus unterschiedlichen Lebensbereichen zum Zeitpunkt des Eintritts in das Programm gesammelt. Die einbezogenen Variablen konzentrieren sich auf das Thema Gesundheit und sind in fünf Klassen eingeteilt: Demographie, Ernährung, Labor, Untersuchung und Fragebogen.

Das Programm zielt darauf ab, repräsentativ für die gesamte nicht-institutionalisierte, d. h. nicht im Gefängnis oder einer geschlossenen Psychiatrie sitzende US-Bevölkerung zu sein. Um dies zu erreichen, wird zusätzlich zu den direkt erhobenen Daten jeder untersuchten Person ein Gewicht zugeordnet zum Ausgleich von Oversampling sowie der Nicht-Beantwortung von Fragen und zur Angleichung an Zensusdaten.

Insgesamt bilden diese beiden Programme zwei extrem reichhaltige Datenquellen, die es erlauben unter der Zuhilfenahme der Datenbank aller Sterbefälle in den USA (Linked Mortality File, LMF), das Mortalitätsrisiko in den USA über eine enorme Anzahl von Risikofaktoren über einen Beobachtungszeitraum von bis zu 25 Jahren zu untersuchen.

Die offizielle Website des National Center for Health Statistics, auf der alle Informationen und Daten gesammelt werden, ist https://www.cdc.gov/nchs/.

3.2. Datenstruktur

Die Struktur der Daten unterscheidet sich von NHANES III zu NHANES Continuous, wobei die zweite Studie komplexer ist.

3.2.1. NHANES III

NHANES III, 1988-1994, enthält Daten für 33.994 Personen im Alter von zwei Monaten und älter, die an der Umfrage teilgenommen haben. Die Daten bestehen aus vier separaten Datenbanken, die dank eines eindeutigen Identifikators

(SEQN) aggregiert werden: Haushalt Erwachsene bzw. Jugendliche (ADULT bzw. YOUTH, Stammdaten), Untersuchung (EXAM) und Labor (LAB).

Die beiden Kohorten, die während des NHANES III-Programms erfasst wurden (1988-1990 und 1991-1994), sind bereits in den Datenbanken einheitlich erfasst.

Darüber hinaus wird anhand des „Linked Mortality File" der Sterblichkeitsstatus für jede erwachsene Person von Studienbeginn bis zur Datenerhebung des Linked Mortality File erfasst:

ADULT	YOUTH	EXAM	LAB	MORT
• SEQN • VarHA_1 • VarHA_2 • …	• SEQN • VarHA_1 • VarHA_2 …	• SEQN • VarHA_1 • VarHA_2 …	• SEQN • VarHA_1 • VarHA_2 …	• SEQN • VarHA_1 • VarHA_2 …

Hier wurde zuletzt das Mortality-File ausgetauscht und es steht nur noch dasjenige aus dem Jahr 2019 zur Verfügung, vergleiche auch die Anmerkungen in 3.4.1.

3.2.2. NHANES Continuous

Die kontinuierliche NHANES-Studie begann 1999. Jedes Jahr werden ca. 7.000 Personen aller Altersgruppen zu Hause befragt; davon absolvieren ca. 5.000 die Gesundheitsuntersuchungskomponente der Umfrage.

Die Struktur von NHANES Continuous ist komplexer als diejenige von NHANES III. Es gibt so viele Tabellen wie Kohorten multipliziert mit der Anzahl der Unterkategorien.

Die Unterkategorien sind Unterabschnitte der fünf Hauptgruppen: Demographie, Ernährung, Labor, Untersuchung und Fragebogen.

Es sind derzeit zehn Kohorten verfügbar, wovon wir acht aufbereiten, da die Sterblichkeitsdaten zum Auswertungszeitpunkt nur bis 2015 veröffentlicht wurden. Die Anzahl der Unterkategorien ist beträchtlich, so dass wir am Ende mehr als 1.200 Tabellen zu aggregieren haben.

Wie bei NHANES III müssen zusätzlich alle Personen mit dem Datensatz zum Sterblichkeitsstatus (Linked Mortality File) verknüpft werden, um ihren Mortalitätsstatus zu erhalten.

Demografie Bis zu 465 Variablen	**Untersuchung** Bis zu 3812 Variablen	**Labor** Bis zu 1432 Variablen	**Fragebogen** Bis zu 2635 Variablen	**Diätetisch** Bis zu 796 Variablen
Alter	BMI	Gesamt-Cholesterin	Haben Sie jemals ge-	Balaststoff gesamt

			raucht?	
Familienstand	Taillenumfang	HDL / LDL	Aktueller Raucher-status?	Zucker gesamt
Verh. Einkommen / Armutsgrenze	Gewicht	Triglyzerid	Alkoholkonsum	Fett gesamt
Gesamtes Familieneinkommen	Körpergröße	Glykohämoglobin	Medizinische Vorgeschichte	Gesamtfett mehrfach ungesättigt
Geburtsland	Kopfumfang	Kotinin	Diabetes	Gesamtfett gesättigt
Geschlecht	Systolischer Blutdruck	Plasma-Nüchtern-Glukose	Krankenversicherungsschutz?	Gesamtfett monoungesättigt
Haben Sie jemals in den Streitkräften der USA gedient	Diastolischer Blutdruck	Folsäure	Beurteilung des allgemeinen Gesundheitszustandes	Um welche Uhrzeit haben Sie mit dem Essen/Trinken begonnen?
Anzahl der Personen in der Familie	Anzahl der Zähne	Anzahl der weißen Blutkörperchen	Schwierigkeitsgrad, um eine Viertelmeile zu gehen	Welche Art von Milch wurde Ihrem Kaffee oder Tee üblicherweise zugesetzt?
Anzahl der Personen im Haushalt	Tragen Sie eine Brille oder Kontaktlinsen zum Lesen oder für den Nahbereich?	Erythrozytenzahl	Schlafstunden	Anzahl, wie oft Barsch in den letzten 30 Tagen gegessen wurde
Niveau/Grad der Ausbildung	Intensität Minuten der Aktivität	Insulin	Kardiovaskuläres Fitnessniveau	Vitamine
Sprache	Steps	Eisen	Minuten der sitzenden Tätigkeit	Energie (Kcal)

Hinweise zur Dokumentation aller verfügbaren Daten:

- Liste der Demografie-Variablen: https://wwwn.cdc.gov/nchs/nhanes/Search/variablelist.aspx?Component=Demographics
- Liste der Untersuchungsvariablen: https://wwwn.cdc.gov/nchs/nhanes/search/variablelist.aspx?Component=Examination
- Liste der Laborvariablen: https://wwwn.cdc.gov/nchs/nhanes/Search/variablelist.aspx?Component=Laboratory
- Liste der Variablen des Fragebogens: https://wwwn.cdc.gov/nchs/nhanes/Search/variablelist.aspx?Component=Questionnaire
- Liste der diätetischen Variablen: https://wwwn.cdc.gov/nchs/nhanes/Search/variablelist.aspx?Component=Dietary
- Liste der Variablen mit beschränktem Zugriff: https://wwwn.cdc.gov/nchs/nhanes/Search/variablelist.aspx?Component=LimitedAccess
  Diese Variablen wie etwa zum Alkoholkonsum von Jugendlichen oder deren psychischer Gesundheit stehen nur über einen speziellen Forschungszugang zur Verfügung und waren nicht Gegenstand unserer Aufbereitung.

3.2.3. Linked Mortality File

Die NHANES-Studienprogramme dienen zur Erhebung des Gesundheitsstatus der US-Bevölkerung zu einem Zeitpunkt als Querschnittsanalyse. Unabhängig von diesen Studien werden alle gemeldeten Todesfälle in den USA erfasst und anhand verfügbarer Merkmale aufbereitet, so dass eine Zuordnung eines Todesfalls zu den Teilnehmern der NHANES-Studien wie auch weiterer Studien der CDC möglich ist.

3.3. Herausforderungen bei der Zusammenführung der Daten

3.3.1. Innerhalb von NHANES Continuous

Homogenisierung: Vorsicht ist geboten, wenn es darum geht, eine gegebene Tabelle über die Kohorten zu aggregieren, da sich die Art und Weise der Erfassung einiger Variablen im Laufe der Zeit ändern kann.

So können sich z. B. der Name einer Variablen ändern, die Einheit der erfassten Größe, die Gruppierung einer Variablen (z. B. bei einem Einkommensband) oder

es können sich weitere Aspekte unterscheiden, die berücksichtigt werden müssen.

Folglich ist eine vorherige Homogenisierungsverarbeitung erforderlich.

Verflachung: Einige Tabellen enthalten mehrere Datensätze für dieselbe Person. Die Anzahl der Schritte wird zum Beispiel sieben Tage lang jede Minute eines Tages aufgezeichnet. Dann gibt es Tausende von Zeilen für eine bestimmte Person.

Um die Verwendung solcher Tabellen zu erleichtern und sie mit den anderen einfacher zu aggregieren, kann es sinnvoll sein, sie zunächst zu glätten.

Für das Beispiel der Anzahl der Schritte können wir z. B. nur die durchschnittliche Anzahl der täglichen Schritte festhalten.

3.3.2. Kombination von NHANES III und NHANES Continuous

Homogenisierung: Einige Informationen können in einer der Studien vorhanden sein, in der anderen jedoch nicht, oder die Informationen können sich von einer Studie zur anderen leicht unterscheiden (z. B. Erfassung der Raucherinformation nur als Ja/Nein, detaillierter als Häufigkeit des Rauchens).

Auch hier, und ebenso bei NHANES Continuous, sind also zunächst einige Vorarbeiten notwendig, um die Informationen über beide Studien anzugleichen.

Gewichtungen: Wie in der Einleitung kurz erwähnt, wird jedem Individuum ein Gewicht zugewiesen, um aus der Studienstichprobe auf die nationale nichtinstitutionalisierte US-Bevölkerung schließen zu können.

Die NHANES-Erhebung gibt klare Regeln vor, wie die Gewichte neu zu berechnen sind, wenn wir mehrere Kohorten über NHANES Continuous kombiniert haben.

Es gibt jedoch keine Anleitung, wie die Gewichte bei der Kombination von NHANES III und NHANES Continuous neu zu berechnen sind.

Da unsere Analysen nicht den Anspruch erheben, eine spezielle Bevölkerung zu beschreiben, haben wir die Gewichte der NHANES III Studie lediglich approximativ unter der Annahme einer konstanten Bevölkerung mit der NHANES Continuous Studie verbunden (Acht NHANES Continuous Studien, jeweils unter sich gewichtet mit dem offiziellen Algorithmus, und dazu die NHANES III Studie, alle neun Studien zu je einem Neuntel). Darüberhinaus haben wir uns auf die Gewichtung der untersuchten Individuen beschränkt – die Menge der lediglich befragten, aber nicht untersuchten Personen ist größer und hätte daher eine andere Gewichtung.

Schlüsselvariablen: Die Personenidentifikation geschieht in beiden Studien sowie den jeweils zugehörigen Sterbedaten über die Variable SEQN. Diese wurde allerdings nicht konsistent von NHANES III nach Continuous durchgeführt, sondern neu begonnen, so dass die Studie selbst als weitere Schlüsselvariable für eine eindeutige Identifikation herangezogen wird.

3.4. NHANES and Linked Mortality File Data-Use-Restrictions

Im Rahmen der NHANES-Studie wird seitens der CDC Zugriff auf "public" und "non-public"-Daten gewährt.

Für die Nutzung der public-Daten gelten folgende Data-Use-Restrictions:

https://www.cdc.gov/nchs/data_access/restrictions.htm

Wesentliche Punkte sind:

- Nutzung nur für statistische Auswertungen und Berichte
- Die Identifizierung einzelner Personen ist untersagt. Hierzu gehört auch die Verlinkung mit weiteren Datenbeständen, die eine Identifizierung einzelner Personen ermöglichen würde.
- Keine Aktivitäten oder Veröffentlichungen zu Methoden, die eine Identifizierung einzelner Personen ermöglichen könnte

3.4.1. Linked Mortality File (LMF)

Für die Auswertungen zur Sterblichkeit werden die NHANES-Daten mit den public-use Linked Mortality Files verknüpft, was auch unter den oben genannten Regeln zulässig ist.

Auch für diese LMF gelten für sich genommen vergleichbaren Regelungen zum Datenschutz, siehe https://www.cdc.gov/nchs/data/datalinkage/public-use-2015-linked-mortality-file-description.pdf

Am 10. Mai 2022 hat die CDC neue Linked-Mortality-Files auf deren Webseite und deren FTP-Server für das Jahr 2019 zur Verfügung gestellt. Unglücklicherweise wurde gleichzeitig dasjenige des Jahres 2015 entfernt, auf dessen Daten dieser Bericht basiert. Die beispielhaften Ergebnisse der Abschnitte 6 und 7 lassen sich damit leider nicht mehr mit den zur Verfügung gestellten Skripten reproduzieren. An der Methodik der Datenaufbereitung hat sich jedoch nichts geändert. Die Skripte können demnach weiterhin ausgeführt werden und produzieren den gewünschten Datensatz.

Durch die Hinzunahme von vier weiteren Jahren an Todesfällen erhöht sich die Anzahl der berücksichtigten Todesfälle deutlich.

4. Unternehmensübergreifende Sterblichkeitsanalysen

Aktuarielle Analysen von unternehmensübergreifenden Portfoliodaten im Bereich der Lebensversicherung sind innerhalb der DAV nicht standardisiert und wurden die letzten Jahre nur ad hoc durchgeführt bei der Erstellung von Reservierungstafeln. In Deutschland gibt es übergreifende Analysen von Portfoliodaten durch einige Rückversicherer. Die Ergebnisse sind allerdings jeweils vertraulich und stehen auch den Mitgliedern der DAV nicht zur Verfügung.

Der NHANES-Datensatz ist ein Beispiel für eine breit angelegte Studie, die mit klassischen und neuen aktuariellen Methoden analysiert werden kann, und die gleichzeitig die Datenschutzbestimmungen der USA und der Datenerheber (CDC) einhält.

4.1. Arbeitsgruppe Big Data in der Lebensversicherung

Nach der Veröffentlichung des stark qualitativen Berichtes „Big Data in der Lebensversicherung" [1], welcher u. a. verschiedenste Anwendungsszenarien von Big Data und ML-Verfahren skizziert, hat die Arbeitsgruppe diese Anwendungsszenarien auf eine praktische Durchführbarkeit geprüft. Ein wesentlicher Prüfaspekt war die Verfügbarkeit von Daten.

Für das Anwendungsszenario der aktuariellen Analyse von biometrischen Daten wurde eine Sichtung öffentlicher Portfoliodaten vorgenommen. Der NHANES-Datensatz stellt dabei sowohl inhaltlich als auch von der Struktur und Organisation den interessantesten, öffentlich verfügbaren Datensatz dar, der Portfoliodaten innerhalb von Versicherungsunternehmen halbwegs ähnlich kommt, und gleichzeitig die Nutzung verschiedenster aktuarieller ML-Verfahren ermöglicht.

Ziel der hier und im weiteren Bericht vorgestellten Sterblichkeitsanalysen ist die Bereitstellung von konkreten Verfahren auf realen Datensätzen. Die parallel zum Ergebnisbericht veröffentlichten Skripte auf github sollten mit überschaubaren Anpassungen auch für andere Portfoliodaten aktuarielle Auswertungen ermöglichen und eine erste Einführung in die benutzten ML-Methoden geben.

Daneben gibt der NHANES-Datensatz wie im Abschnitt 3 vorgestellt ein sehr reichhaltiges Anschauungsprogramm und Analysemöglichkeiten von Einflussfaktoren auf die Sterblichkeit aufgrund der umfangreichen Variablenauswahl. Deshalb ist die Bereitstellung des aufbereiteten NHANES-Datensatzes bereits an sich wertvoll und sollte Quelle für weitere Untersuchungen sein. Diese können sich sowohl auf inhaltliche Themen fokussieren unter Berücksichtigung, dass die Daten aus den USA stammen, als auch auf die Anwendung bisher wenig oder ungenutzter ML-Verfahren in der aktuariellen Analyse.

4.2. Arbeitsorganisation

Nach der Festlegung des neuen Projektzieles durch die Arbeitsgruppe wurden voneinander unabhängige Arbeitsbereiche festgelegt, die Kleingruppen der Arbeitsgruppe bearbeitet haben:

- Datenaufbereitung

- Exploration und Imputation
- Klassische Sterblichkeitsanalyse
- Regressionsverfahren
- Survival-Analysis

Die ersten beiden Punkte werden in diesem ersten Ergebnisbericht behandelt. Die weiteren Analysen folgen dann im vollständigen Ergebnisbericht.

Von technischer Seite wurden folgende Plattformen verwendet:

- Sharepointserver der DAV für Arbeitsgruppen: Die Sharepointseite wurde im Wesentlichen für einen Dateiaustausch benutzt, insbesondere Worddokumente, Dokumentation des NHANES-Datensatzes, und ähnliches.
- Github.com: Für lauffähige Skripte zur Datenaufbereitung und Datenanalyse wurde ein privater github-Account verwendet. Die Verzeichnisstruktur lehnt sich an obige Gruppen an. Die parallele Bearbeitung von Skripten und das einheitliche Zusammenführen der Skripte war die Hauptanwendung der git-Funktionalitäten.
- R-Studio: Die Mehrzahl der Arbeitsgruppenmitglieder nutzt R als Skriptsprache für die Datenaufbereitung und Analyse. R-Studio inkl. Github-Funktionalität und R-Markdown mit einer Vielzahl von R-Paketen wurden verwendet, vgl. auch die Skripte im DAV-Github-Projektverzeichnis.
- Python: Ein Teil der Arbeitsgruppe hat Python, Python-Pakete und jupyter-Notebooks verwendet, insbesondere für die Datenexploration. Diese sind entsprechend auf github abgelegt.
- Excel: Auch mit Excel sind aufgrund der überschaubaren Größe des NHANES-Datensatzes von knapp 65.000 Datensätzen einfache Strukturanalysen möglich.

Bei ungefähr sieben aktiven Contributors der Python- und R-Skripte sind git und github eine wertvolle Plattform, um eine mittelgroße Anzahl von Skripten konsistent und lauffähig über alle Arbeitsgruppenmitglieder zu halten.

Schwierigkeiten haben sich im Lauf des Projekts eher im Bereich des Austauschs der jeweils letzten Version der aufbereiteten Daten ergeben.

Zunächst zum Datenformat des Austauschs: Nicht alle Arbeitsgruppenmitglieder, die in Python gearbeitet haben, konnten frei Pakete installieren, die ihnen Zugriff auf das native R-Datenformat gegeben hätten, so dass wir letztlich auf CSV Dateien zurückgegriffen haben.

Außerdem aufgrund der langen Laufzeit der Skripte zur Datenaufbereitung: Hierdurch war es nicht effizient, dass sich die AG-Mitglieder nach erfolgter Aktualisierung des Herleitungscodes die Daten immer neu erzeugen.

Hier hat sich auch gezeigt, dass die zu Beginn postulierte „Unabhängigkeit" der Tätigkeiten nicht vollständig gegeben war. Praktisch hat dies dazu geführt, dass

in github sogenannte Branches (Entwicklungszweige) des Projekts angelegt wurden: Weiterentwicklungen der Datenherleitung wurden in einem Nebenzweig ausgeführt und erst nach Abstimmung mit der Arbeitsgruppe in den Hauptzweig überführt, in dem verschiedenen Analyseskripte weiterentwickelt wurden, die dann eher voneinander unabhängig waren.

5. Datenaufbereitung

Eine zentrale Herausforderung des Projektes war die Datenaufbereitung. Die NHANES-Studie stellt ihre Daten verteilt auf eine Vielzahl einzelner Tabellen zur Verfügung. Diese galt es im ersten Schritt zusammenzuführen und zu harmonisieren – siehe Abschnitt 3.2.

Zu Beginn der Analysen wurde ein bereits fertig aufbereiteter Datensatz verwendet, der der Arbeitsgruppe von der SCOR für die Zwecke der Analyse zur Verfügung gestellt wurde. Der Abschnitt 6.1 zur Exploration eines komplexen Datensatzes baute ursprünglich auf diesem Datensatz auf.

Zum Zwecke der Veröffentlichung haben wir uns in der AG jedoch entschlossen, einen eigenen Aufbereitungscode zu schreiben und mit diesem parallel einen zweiten Datensatz zu erstellen, der sich am ursprünglichen Datensatz der SCOR orientiert. Sowohl Aufbereitungscode als auch den zugehörigen Datensatz möchten wir veröffentlichen, damit interessierten Kollegen eine Datenbasis zu Verfügung steht, um eigene, ggf. weiterführende Untersuchungen durchzuführen und die Ergebnisse unserer Analysen nachvollziehen zu können.

Um die bereits bestehenden Analysen möglichst nahtlos weiter verwenden zu können, haben wir uns in der Benennung und Belegung der Variablen an die Konventionen des ursprünglichen Datensatzes der SCOR gehalten. Eine Vielzahl der ursprünglichen Auffälligkeiten, die während der explorativen Datenanalyse identifiziert wurden (Inkonsistenzen in den Belegungen, unvollständige Behandlung spezieller Fehlerwerte in kontinuierlichen Variablen, etc.), wurden im neuen Datensatz direkt behoben.

Unsere Aufbereitung besteht im Wesentlichen aus den folgenden Schritten, organisiert in einem R-Projekt, welches wir über das DAV-Konto auf GitHub zur Verfügung stellen:

- Herunterladen der NHANES III Daten von der CDC-Website und Aufbereiten, d. h. Umbenennung und Umbelegung im Hinblick auf Konsistenz mit NHANES Continuous
- Herunterladen der NHANES Continuous Daten durch das nhanesA-Paket und Aufbereiten, d.h. Umbenennung und Umbelegung im Hinblick auf Konsistenz mit NHANES Continuous
- Herunterladen der 2015 (neu: 2019) Sterbedaten Linked Mortality File von der CDC Website und Aufbereiten
- Herunterladen der Aktivitätsdaten von der CDC Website und Aufbereiten, insbes. Aggregation auf die Durchschnittswerte pro Individuum
- Zusammenführen der Datensätze und Erzeugung von Zusatzfeldern zur einfacheren Analyse der Daten im Hinblick auf Sterblichkeitsanalysen

Die Organisation des Projekts fasst alle Einstellungen zur Liste der erwünschten Studienjahre und -variablen in einem Initialisierungsskript zusammen.

Zusätzliche Variablen oder zukünftige Jahre müssen hier und in den speziellen Aufbereitungsschritten ergänzend aufgenommen werden. Für jede neue Variable ist ein einheitlicher Name und Belegungsplan notwendig, der in den Skripten zu NHANES III und Continuous einzuarbeiten ist. Für neue Jahre ist zu prüfen, ob sich an den verwendeten Variablen gegebenenfalls etwas geändert hat.

Letztlich ergeben sich für den Datensatz NHANES_ALL folgende Variablen und Bedeutungen:

Variablenname	Beschreibung	Variable NHANES III	Variable NHANES Continuous
SEQN	Personen ID des CDC in NHANES III und NHANES Continuous, getrennt für die beiden Studien, d. h. für eine eindeutige Identifizierung einer Person ist die Angabe Studie und SEQN erforderlich.	SEQN	SEQN
ID	Eigene globale Personen ID		
Wave	Zyklus der NHANES Continuous Studie (A-H)		
Phase	Zyklus der NHANES III Studie (1-2)	SDPPHASE	
Cycle	Eigener globaler Zyklus (1,2,A-H)		
Study	Eigene ID der Studie	3/NHANES III	4/NHANES Continuous
SampleWeightMEC	Stichprobengewicht bzgl. Untersuchter Personen	WTPFEX6	WTMEC4YR(1999->2002)/MEC2YR(2003->2014)
YearDOS	Jahr des Eintritts in die Studie		
Gender	Geschlecht	HSSEX	RIAGENDER
AgeDOS	Alter bei Eintritt in die Studie	HSAGEIR	RIDAGEYR
Tot_Income_family		HFF19R	INDFMINC (1999->2006) INDFMIN2 (2007->2014)
Tot_Income_household		#N/A	INDHHINC (1999->2006) INDHHIN2 (2007->2014)
Educational_Level_20plus		#N/A	DMDEDUC2
Marital_status		HFA12	DMDMARTL
EverSmoker		HAR1	SMQ020
CurrentSmoker		HAR3	SMQ040
highBP		HAE2	BPQ020
highChol		HAE7	BPQ080
Alcohol		MAPE7	ALQ150 (1999-2010) ALQ151 (2011-2014)
Diabetes		HAD1	DIQ010
pastHeartAtt		HAF10	MCQ160E
pastStroke		HAC1D	MCQ160F
pastHeartDis		#N/A	MCQ160C
pastHeartFailure		#N/A	MCQ160B
pastAngina		#N/A	MCQ160D
pastCancer		#N/A	MCQ220

Industry		#N/A	#N/A
Occupation		#N/A	#N/A
Health_insurance_coverage		#N/A	HID010 (1999-2004) HIQ011 (2005-2014)
Cover_Private_insurance		HFB10	HID030A (1999-2004) HIQ031A (2005-2014)
Cover_Medicare		HFB1	HID030B (1999-2004) HIQ031B (2005-2014)
Cover_Medicaid_CHIP		HFB6	HID030C (1999-2004) HIQ031D (2005-2014)
Cover_other_government_insurance		#N/A	HID030D (1999-2004) HIQ031I (2005-2014)
Cover_any_single_plan		#N/A	HID030E (1999-2004) HIQ031J (2005-2014)
compare_activity_same_age		HAT28	PAQ520 (1999-2006 only)
difficulty_walk_quarter_mile		HAH1	PFQ061B
Cardiovascular_fitness_level		#N/A	CVDFITLV
Predicted_VO2MAX		#N/A	CVDVOMAX
Estimated_VO2MAX		#N/A	CVDESVO2
BMI		BMPBMI	BMXBMI
Waist_circumference		BMPWAIST	BMXWAIST
Weight_body_metric		BMPWT	BMXWT
Standing_height		BMPHT	BMXHT
total_fat_sat		DRPNSFAT	DRXISFAT (1999-2002) DR1ISFAT (2003-2014) DR2ISFAT (2003-2014)
total_fat_monounsat		DRPNMFAT	#N/A
total_fat_polyunsat		DRPNPFAT	DRXIPFAT (1999-2002) DR1IPFAT (2003-2014) DR2IPFAT (2003-2014)
total_fat		DRPNTFAT	DRXITFAT (1999-2002) DR1ITFAT (2003-2014) DR2ITFAT (2003-2014)
total_fiber		DRPNFIBE	DRXIFIBE (1999-2002) DR1IFIBE (2003-2014) DR2IFIBE (2003-2014)
TotalCholesterol		TCPSI	LBDTCSI
HDL		HDPSI	LBDHDLSI (1999-2002) LBDHDDSI (2003-2014)
LDL		LCPSI	LBDLDLSI
Triglyceride		TGPSI	LBDTRSI
Glycohemoglobin		GHP	#N/A
plasGluMG		G1P	LBXGLU

plasGluMMOL		G1PSI	LBXGLUSI (1999-2002) LBDGLUSI (2003-2014)
Cotinine		COP	LBXCOT
pastHeartFail		HAC1C	#N/A
pastSkinCancer		HAC1N	#N/A
pastOtherCancer		HAC1O	#N/A
Family_Poverty_income_ratio		DMPPIR	INDFMPIR
Tot_Income_family_comp20000		HFF18	#N/A
general_health_condition		HAB1	HSD010 (except 1999-2000)
Pulse		HAZA5	#N/A
SBP		PEP6G1	BPXSY1
DBP		PEP6G3	BPXDI1
Smoker		0	0

5.1. Ablauflogik der Datenaufbereitung

Auf dem Github-Repository der Arbeitsgruppe liegen alle relevanten Dateien im Verzeichnis NHANES_ORIGINAL. Die Tabelle 2 gibt eine Übersicht der verwendeten R-Skripte. Die heruntergeladenen Dateien werden im Verzeichnis „./data/" gespeichert. Es werden folgende R-Libraries verwendet: tidyverse, nhanesA (Einlesen von NHANES Continuous-Tabellen), arsenal (Funktionen zum Datenabgleich) und haven (Einlesen von SAS-Daten). Lauffähige und getestete Versionen von Betriebssystem, R-Version und R-Skripten finden sich in der Tabelle 1.

Je nach persönlicher Konfiguration und Nutzung von Paketmanagern sollten die nötigen R-Pakete im Vorfeld installiert werden. Die Skripte testen auf das Vorhandensein der R-Pakete. Falls Pakete nicht verfügbar sind, werden diese in die aktuelle R-Umgebung installiert.

Probleme beim Herunterladen der Daten werden häufiger durch IT-Sicherheitseinrichtungen in Konzernen und Versicherungen wie VPN, Firewalls etc. verursacht. Direkte Fehlermeldungen werden anscheinend nicht zwingend angezeigt. Falls das Herunterladen der Daten auffällig lange, d. h. länger als 2 Stunden dauert, sollten die Skripte in einer anderen Umgebung ausgeführt werden. Geeignet sind insbesondere anscheinend Heim-PCs.

5.1.1. Lauffähige Umgebungen

Folgende Kombinationen wurden getestet:

Tabelle 1: Umgebungen, in denen die Skripte getestet wurden und lauffähig waren

Betriebssystem	Windows 10	Ubuntu 20.04 LTS		
R-Version	4.0.2 (64 bit)	3.6.3 (64-bit)		
tidyverse	1.3.1	1.3.0		

nhanesA	0.6.5.3	0.6.5.3		
arsenal	3.6.3	3.6.2		
haven	2.4.3	2.3.1		

5.1.2. Beschreibung der R-Skripte

Tabelle 2: R-Skripte zur Erstellung des Datensatzes NHANES_ALL

NHANES_full_run.R	Initialisierung, Löschen aller evtl. vorhandenen Daten aus vorhergehenden Läufen, Löschen aller heruntergeladenen Daten in „./data/".
NHANES_main.R	Zentrale Ablaufdatei. Es werden nacheinander die R-Dateien ausgeführt: 1. NHANES_init.R 2. NHANES_Cont.R 3. NHANES_III_proc.R 4. NHANES_mort.R 5. NHANES_Activity.R Die in verschiedenen dataframes eingelesenen und verarbeiteten Daten werden anschließend zusammengeführt und weitere erst nach der Aggregierung erzeugbare Variablen für die Sterblichkeitsanalyse erstellt. Ausgabe des finalen Datensatzes in NHANES_ALL.RData und NHANES_ALL.csv
NHANES_init.R	Initialisierung und Festlegung der einzelnen aufzubereitenden Studien aus NHANES III und NHANES Continuous. Festlegung des Enddatums der Beobachtungen der Sterblichkeitsdaten, welches nun neu auf 2019 gesetzt wurde. Definition der Tabellen der NHANES Continuous-Studie, die heruntergeladen werden sollen. Definition der Namen der Outputvariablen des finalen Datensatzes.
NHANES_Cont.R	Herunterladen der NHANES Continuous-Tabellen für die festgelegten Jahre. Umbennen und Umkodieren der Zielvariablen, hierbei erfolgt teilweise auch eine Fehlerkorrektur. Ausgabe Daten in /data/NHANES_CONT.RData
NHANES_III_proc.R	Die vom Skript NHANES_III.R (vgl. folgende Zeile) bereits eingelesenen Daten der NHANES III-Studie werden aufbereitet. Hierzu erfolgt wie für die NHANES Continuous-Daten

	eine Umbenennung der Variablennamen und eine Umkodierung von Variablen für weitere, einheitliche Auswertungen.
NHANES_III.R	Herunterladen der NHANES III-Daten vom CDC. Die Daten liegen in verschiedenen SAS-fixed-format-Dateien vor. Es werden die Gruppen Adult, Exam und Lab heruntergeladen.
NHANES_Activity.R	Die „Activity"-Daten enthalten detaillierte stichprobenhafte Informationen über die sportliche Aktivität von Personen. Für die weitere Analysen werden direkt Durchschnittswerte pro Beobachtungstag und pro Person gebildet.
NHANES_mort.R	Die Mortality-Daten (Linked Mortality File) gehören nicht direkt zum NHANES-Studienprogramm. Anhand der Todesfallstatistik der USA werden Todesfalldaten erzeugt, die den verschiedenen Studienprogramm der CDC zugeordnet werden können. Die Verbindung erfolgt über die SEQ-Number (SEQN). Der Datensatz enthält eher wenige weitere Variablen. Zuletzt die angesprochene Änderung des Datensatzes bis zum Jahr 2019 für die Todesfalldaten.

6. Methoden und Algorithmen

6.1. Exploration

Eines der ersten Projekte der AG bestand - wie es bei Data-Science-Projekten üblich ist - in der explorativen Datenanalyse. Diese Datenanalyse wurde auf Basis des ursprünglichen Datensatzes der SCOR durchgeführt und später für die neu von der AG hergeleiteten Daten aktualisiert. Dafür wurden die Daten zuerst grob hinsichtlich Anzahl der Datensätze, verfügbarer Merkmale und fehlender Werte analysiert. Weitere Analysen konzentrierten sich auf die Zielvariable "Sterbealter" und ihre Abhängigkeiten zu den anderen Variablen. Abschließend wurden die einzelnen Merkmale in Bezug auf ihre Ausprägungen und auf mögliche Ausreißer genauer untersucht. Da die explorative Datenanalyse beliebig detailliert vorgenommen werden kann und hier lediglich ihre Grundzüge vorgestellt werden sollten, geben wir an geeigneten Stellen Hinweise für mögliche tiefergehende Analysen.

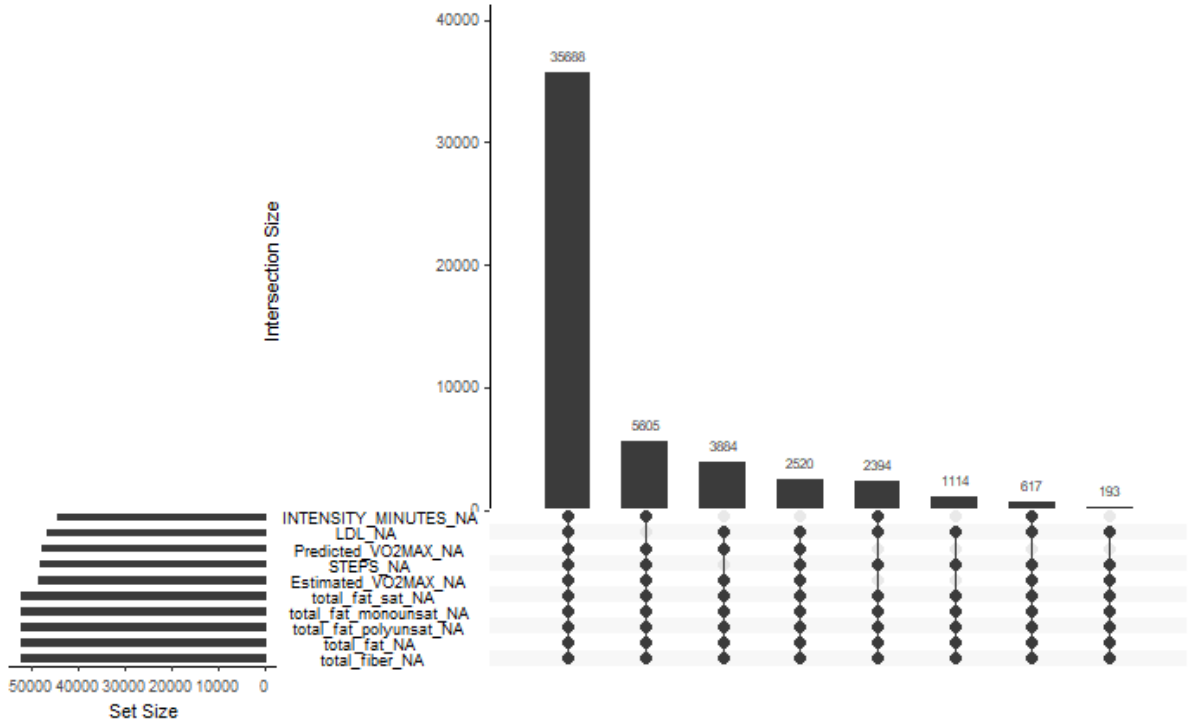
Grobanalyse / Fehlende Daten

Der aggregierte NHANES-Datensatz besitzt ca. 65.000 Beobachtungen und 77 Merkmale. Das sind vergleichsweise viele Informationen und damit kam der Vorauswahl von Merkmalen mit relevantem Erklärungsgehalt für die weiterführenden Modelle eine wichtige Rolle zu. Eine sehr simple Prüfung war die Identifikation von Variablen mit nur einer Ausprägung. Gemäß der Analyse trifft dies nur auf die Variable „eligstat" zu (Angabe, ob ein Datensatz Sterblichkeitsdaten aufgrund von Alter und anderen Vertraulichkeitskriterien liefern kann – hierauf wurde schon im Vorfeld gefiltert). Diese liefert bei der Modellierung keinen Informationsgehalt und konnte daher im weiteren Verlauf der Arbeiten ignoriert werden. Als Ausbaustufe dazu könnten auch Merkmale identifiziert werden, die abgesehen von wenigen Datensätzen immer die gleiche Ausprägung haben.

Im nächsten Schritt wurde ermittelt, wie gut die einzelnen Merkmale befüllt wurden. Die Analyse zeigte, dass viele Merkmale schlecht befüllt waren. Bei fünf Variablen lagen überhaupt keine Informationen vor. Bei weiteren 30 Merkmalen waren weniger als die Hälfte der Beobachtungen befüllt. Dabei muss unterschieden werden zwischen einem nicht belegten Feld und der bewussten Belegung eines Merkmals mit der Ausprägung "Information nicht verfügbar". Die zweitgenannten Fälle wurden bei der Analyse nicht berücksichtigt. Durch die teils sehr lückenhafte Befüllung einiger Merkmale hat sich die AG entschlossen, Imputationsverfahren zur Schätzung der fehlenden Informationen auszuprobieren, vgl. auch den Abschnitt zur Imputation 6.2.

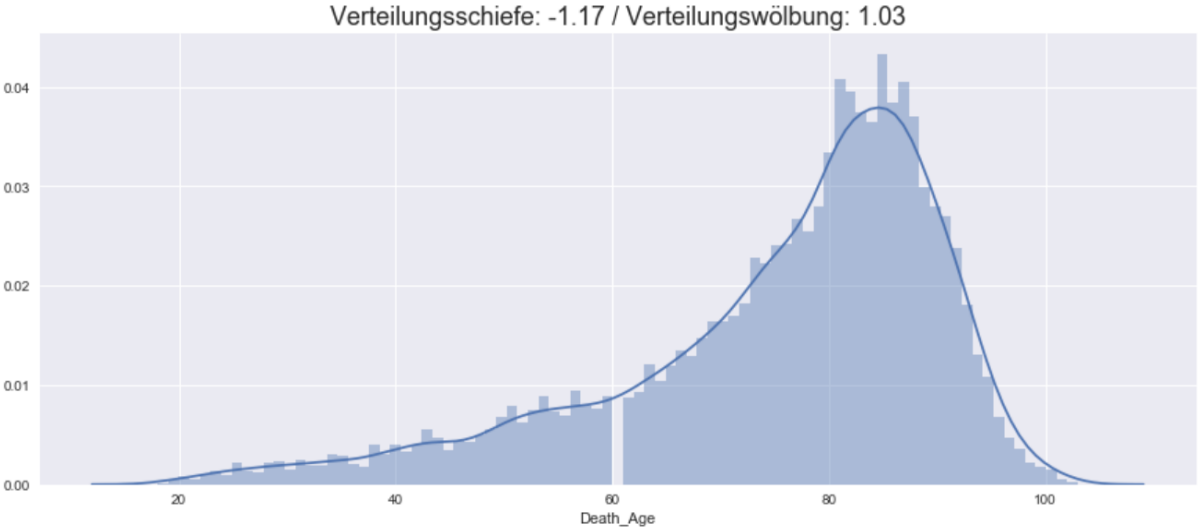
Weiterführende Analysen zu fehlenden Daten könnten deren Struktur und gegenseitige Abhängigkeit untersuchen. Gibt es beispielsweise Merkmale, die häufig zusammen fehlen? Oder ist ein Merkmal häufig dann befüllt, wenn ein anderes Merkmal nicht befüllt wurde? Solche Analysen können weitere Hinweise für die

Datenaufbereitung und die Modellierung geben, wurden von der AG jedoch nur kurz angeschnitten, wie in der nächsten Abbildung zu sehen.



Zielvariable "Sterbealter"

Viele klassische und Machine-Learning-Modelle liefern eine bessere Vorhersagegüte, wenn die Zielvariable normalverteilt ist. Daher haben wir die Verteilung der Sterbealter analysiert und festgestellt, dass diese linksschief ist.



Vor der Anwendung von Machine-Learning-Verfahren könnte das Sterbealter daher z. B. durch

$$\ln\left(\max(Sterbealter) + 1 - Sterbealter\right)$$

transformiert werden, damit die Zielvariable eher einer Normalverteilung folgt.

Im zweiten Schritt wurde die Korrelation zwischen dem Sterbealter und den anderen Merkmalen untersucht, um aussagekräftige Merkmale zu identifizieren. Die Analyse beschränkte sich dabei auf numerische Variablen. Im Ergebnis konnten die Merkmale einer von vier Kategorien zugeordnet werden:

- Starke Korrelation, altersbezogenes Merkmal (z. B. "yearDOB" / Geburtsjahr)
- Starke Korrelation, aber nur bei einem geringen Anteil der Datensätze befüllt (z. B. "Steps" / Anzahl Schritte)
- Starke Korrelation, vielversprechendes Merkmal für Modellierung (z. B. "Standing_Height" / Größe der Person)
- Schwache Korrelation (z. B. "Waist_cirumference" / Hüftumfang)

Diese Analyse gab uns wichtige Hinweise für die Auswahl geeigneter Merkmale zur Modellierung.

Detailanalyse - Häufigkeiten

In der Detailanalyse haben wir in einem ersten Schritt für alle Merkmale die Häufigkeit der verschiedenen Ausprägungen untersucht. Dabei fielen ganz unterschiedliche Punkte auf, die dann im Rahmen der Aufbereitung für den von der AG hergeleiteten Datensatz adressiert werden konnten.

Bei einigen numerischen Variablen traten häufig auffällige Werte auf, z. B. die Belegung "88888" bei "Waist_circumference" oder "888888" bei "Family_Poverty_income_ratio". Laut Dokumentation stehen diese Belegungen dafür, dass die Angaben nicht vorhanden sind. Wir haben dies berücksichtigt, indem neue binäre Variablen generiert wurden, die angeben, ob das jeweilige Merkmal belegt ist oder nicht.

Bei der Variablen "Family_Poverty_income_ratio" fiel zusätzlich die Ausprägung "5" auf. Der Wert stach heraus, da er einerseits die häufigste Ausprägung war und andererseits eine vergleichsweise glatte Zahl war. Die übrigen Ausprägungen hatten überwiegend Nachkommastellen, so wie es bei einer Quote auch zu erwarten ist. Laut Dokumentation steht "5" eigentlich für "5 oder größer". Diese Information könnte man ebenfalls in einer gesonderten Variablen festhalten. Denn für einen Machine-Learning-Algorithmus haben z. B. die Ausprägungen 5 und 4,5 den gleichen metrischen Abstand wie 4,5 und 4. Doch mit dem Zusatzwissen, dass die Ausprägung bei 5 gekappt wurde, ist klar, dass der Abstand zwischen 5 und 4,5 anders gewertet werden sollte als der Abstand zwischen 4,5 und 4.

Die Analyse zeigte auch diverse kategoriale Variablen, die neben einer binären Klassifikation in "ja" und "nein" zusätzlich Zahlenwerte als Ausprägung enthielt. So gab es z. B. für die Variable "Cover_Medicare" folgende Ausprägungen:

- No

- Yes
- 15.0
- 15

Auch hier brachte ein Blick in die Dokumentation Licht ins Dunkel. Die Werte "15" und "15.0" waren ebenfalls Schlüssel mit der Bedeutung "Yes", die wir entsprechend ersetzt haben. Bei den anderen Variablen mit ähnlicher Ausgangssituation sind wir analog vorgegangen.

Die größte Herausforderung stellte die Schlüsselung der Einkommensvariablen "Tot_Income_family" und "Tot_Income_household" dar. Während der verschiedenen Erhebungen der NHANES-Studie kamen bei diesen Variablen unterschiedliche Schlüsselungen zum Einsatz, die nicht ohne weiteres zusammengeführt werden konnten. So gab es anfangs nur die Unterscheidung zwischen Einkommen ober- bzw. unterhalb von 20.000 USD. Später wurde das Einkommen in feineren Abstufungen abgefragt. Zudem wurde die abgefragte Höchstkategorie im Zeitverlauf von "mehr als 50.000 USD" über "mehr als 75.000 USD" hin zu "mehr als 100.000 USD" angepasst.

Wir haben zwei Möglichkeiten ausprobiert, um diese Merkmale zu schlüsseln. In der ersten Variante haben wir das Einkommen als kategoriales Merkmal behandelt. Dafür wurden lediglich unterschiedliche Schlüssel mit eigentlich gleicher Bedeutung zusammengefasst. So wurde in den NHANES-Daten z. B. der Einkommensbereich von 0 USD bis 4.999 USD teilweise ausgeschrieben ("$0 to $4999") und teilweise durch die Kategorie "1" repräsentiert. Im Ergebnis hatten wir eine Schlüsselung des Einkommens durch 17 Kategorien, die sich jedoch teilweise in den zugrundeliegenden Einkommensbereichen überschnitten.

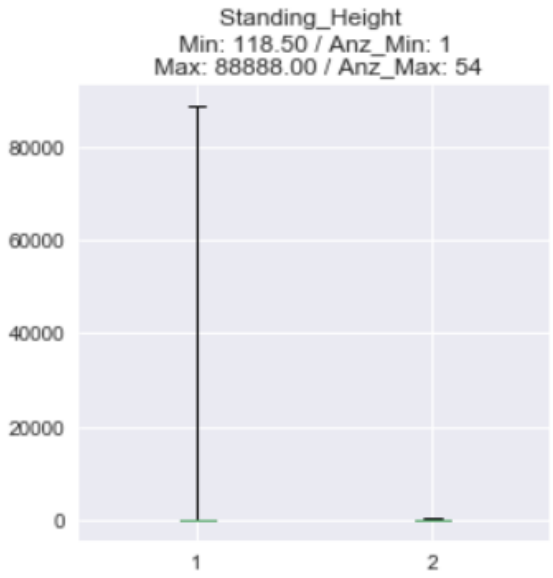
Als zweite Variante haben wir daher eine Aufbereitung als numerisches Merkmal versucht. Damit können auch Metrik und Ordnungsrelationen zwischen den Einkommensklassen abgebildet werden. Dafür wurde jedem Einkommensintervall der jeweilige Mittelwert zugewiesen. Die Kategorie "5.000 USD bis 9.999 USD" wurde in dieser zweiten Variante also z. B. durch den Wert 7.500 repräsentiert. Für halboffene Intervalle, wie z. B. "über 75.000 USD" haben wir eine fiktive Einkommensobergrenze von 225.000 USD angenommen, um den Mittelwert des Intervalls bestimmen zu können. Diese Obergrenze wurde aus dem 95%-Quantil einer amerikanischen Einkommensstatistik geschätzt.

Detailanalyse - Ausreißer

Den Abschluss der explorativen Datenanalyse bildete eine Untersuchung der numerischen Variablen bzgl. möglicher Ausreißer. Dafür haben wir für jede Variable zwei Boxplots erstellt. Während der erste Boxplot alle Werte darstellen sollte, wurden für den zweiten Boxplot das Minimum und das Maximum entfernt. Auf diese Weise konnten wir abschätzen, ob die Ausreißer alle auf einen bestimmten

Wert entfallen oder ob es noch weitere Ausreißer in einer ähnlichen Größenordnung gibt.

Diese Analyse lieferte uns weitere Hinweise für Variablenbelegungen, die als "Angabe nicht verfügbar" interpretiert werden sollten. So zeigte z. B. die Grafik für die Variable Standing_Height den Wert "88888" als Ausreißer.

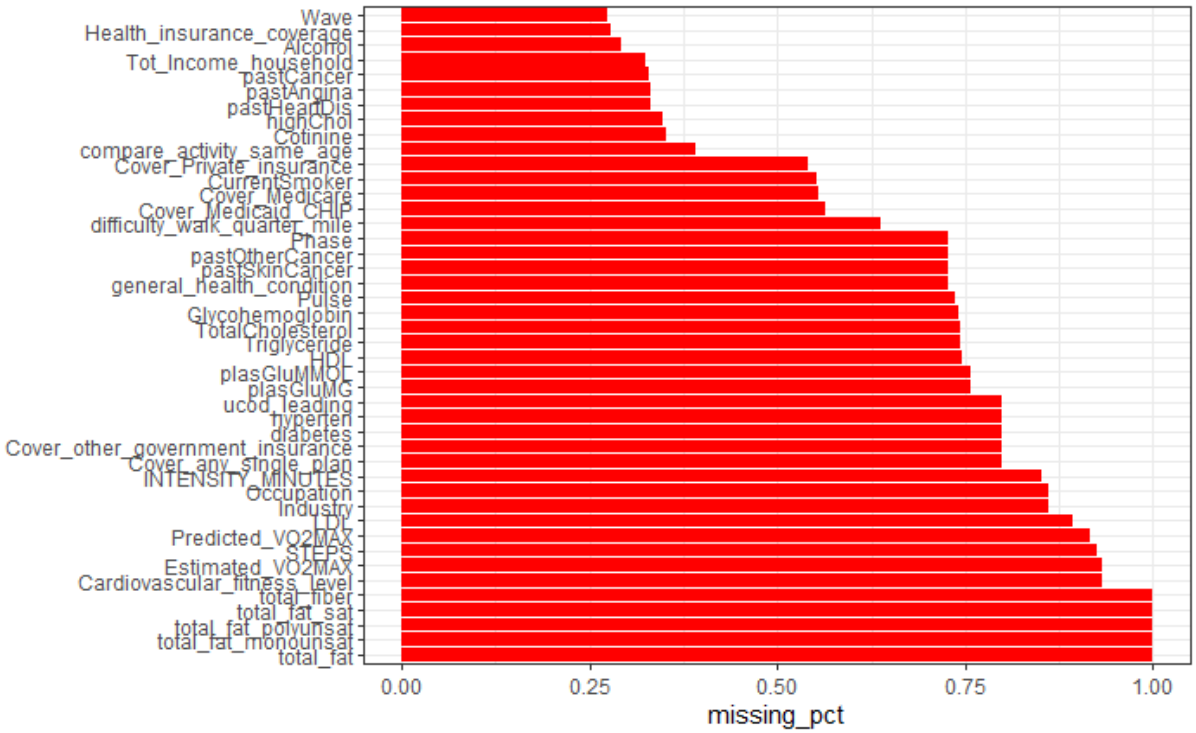


Da dieser Wert im Datensatz jedoch nur 54-mal vorkam, fiel die Sonderbelegung bei der Häufigkeitsanalyse nicht auf.

Auch bei anderen Variablen konnten Ausreißer identifiziert werden, die bei einem realen Projekt adressiert werden müssten, um die Modellierung der Zielvariablen nicht zu verzerren.

6.2. Imputation

Fehlende Daten gibt es in vielen Studien, so auch im NHANES Datensatz. Bei 31 Merkmalen ist weniger als die Hälfte der Beobachtungen befüllt, bei vielen weiteren Merkmalen fehlt mehr als 25% der Beobachtungen.



Der Umgang mit fehlenden Daten ist bei der Vorverarbeitung des NHANES Datensatzes insofern wichtig, da viele Algorithmen für maschinelles Lernen keine fehlenden Werte unterstützen. Dabei gibt es unterschiedliche Vorgehensweisen mit fehlenden Werten umzugehen, wie beispielweise:

- Löschen der Zeilen mit fehlenden Werten
- Single Imputation wie das Ersetzen mit Mean oder Median
- Verwendung von Algorithmen, die mit fehlenden Werten umgehen können (z.B. naive Bayes oder der k-Nearest-Neihbors Algorithmus, die jedoch leider beispielweise im scikit-learn Paket in Python nicht in einer Art und Weise implementiert sind, als dass sie mit fehlenden Werten umgehen könnten)
- Multiple Imputation mit verschiedenen R-Paketen (MICE; Amelia), worauf wir uns in der AG konzentriert haben

Im MICE Paket ist multiple Imputation unter Verwendung der Fully Conditional Specification (FCS) implementiert, während das Paket Amelia Imputation mithilfe des Bootstrap-EM-Algorithmus ausführt.

Auch wenn multiple Imputation in der Theorie durchaus geeignet ist auch eine große Menge an fehlenden Werten zu ersetzen, sind wir in der Praxis schnell an Grenzen gestoßen. Bei einer großen Menge fehlender Daten ist eine höhere An-

zahl an Imputationen sinnvoll. Das hat jedoch zu einer Laufzeit von mehreren Stunden geführt, was für Demo-Zwecke weniger geeignet ist.

Ein weiteres Problem im NHANES-Datensatz war das Vorhandensein von Kollinearität. Einige Pakete können damit umgehen, da hier kollineare Variablen automatisch entfernt werden können (wie beispielsweise beim Paket MICE). Bei anderen Paketen muss man diese Variablen jedoch manuell vor einer Imputation entfernen.

Was noch nicht in der AG betrachtet wurde, ist zum einen ein Vergleich der Imputationsverfahren, wie auch eine Überprüfung der Güte der Imputation. Beides ist nötig, um die imputierten Daten sinnvollerweise für spätere machine learning Methoden zu verwenden.

6.3. Beschreibung der Datenanalyse-Skripte

Tabelle 3: Skripte zur Analyse des Datensatzes NHANES_ALL

NHANES-EDA.ipynb	Beispielhafte explorative Datenanalyse: Analyse des Datensatzes NHANES_ALL hinsichtlich verfügbarer Merkmale, fehlender Daten, Abhängigkeiten anderer Variablen zur Zielvariablen und der Suche nach Ausreißern.

7. Ausblick

Der folgende Abschnitt gibt einen ersten kleinen Ausblick auf die Auswertungen auf der Grundlage der aufbereiteten NHANES-Daten. Eine tiefergehende Analyse mittels verschiedener Methoden wird im zweiten Teil dieses Berichtes vorgestellt. Durch die veränderte Datengrundlage des Todesfalldaten für die NHANES III-Studie ergeben sich mit diesen aktuell zur Verfügung gestellten Daten abweichende Ergebnisse.

7.1. Klassische Analyse

Die nach den vorhergehenden Abschnitten aufbereiteten und einer ersten explorativen Datenanalyse unterzogenen NHANES-Daten werden für eine klassische Analyse der Sterblichkeit weiter aufbereitet. Hierfür wird für die Anwendung der Actuarial Exposure Method für jeden Datensatz, der im Endeffekt eine Beobachtung vom Eintritt in die Studie bis zum Enddatum der Studie oder eines möglichen Todesfalls ist, die Zeit unter Risiko in Abhängigkeit vom Alter bestimmt.

Aus jeder einzelnen Beobachtung in den NHANES-Daten ergibt sich damit eine Anzahl neuer Datensätze, die die Entwicklung einer Person über alle erreichten Alter widerspiegeln. Die weiteren Co-Variablen werden unverändert übernommen. Hierdurch vergrößert sich das Datenvolumen von ca. 65.000 Datensätzen auf ca. 735.000 Datensätze.

Für eine erste klassische Analyse sind die betrachteten Variablen:

- Geschlecht
- Raucherstatus
- Alkoholmissbrauch
- Familienstatus
- Familieneinkommen
- Höchster Ausbildungsabschluss

Um Actual vs. Expected-Ergebnisse zu bestimmen, wird in einem ersten Schritt eine Referenztafel aus dem vorliegenden Datensatz bestimmt.

7.1.1. Referenztafel

Für die Herleitung der Referenztafel werden die rohen Sterbewahrscheinlichkeiten getrennt nach Männern und Frauen mit dem Verfahren von Whittaker-Henderson ausgeglichen.

Abbildung 1: Referenztafel der Männer aus den vorliegenden Datensätzen hergeleitet, logarithmierte Sterblichkeit pro Einzelalter

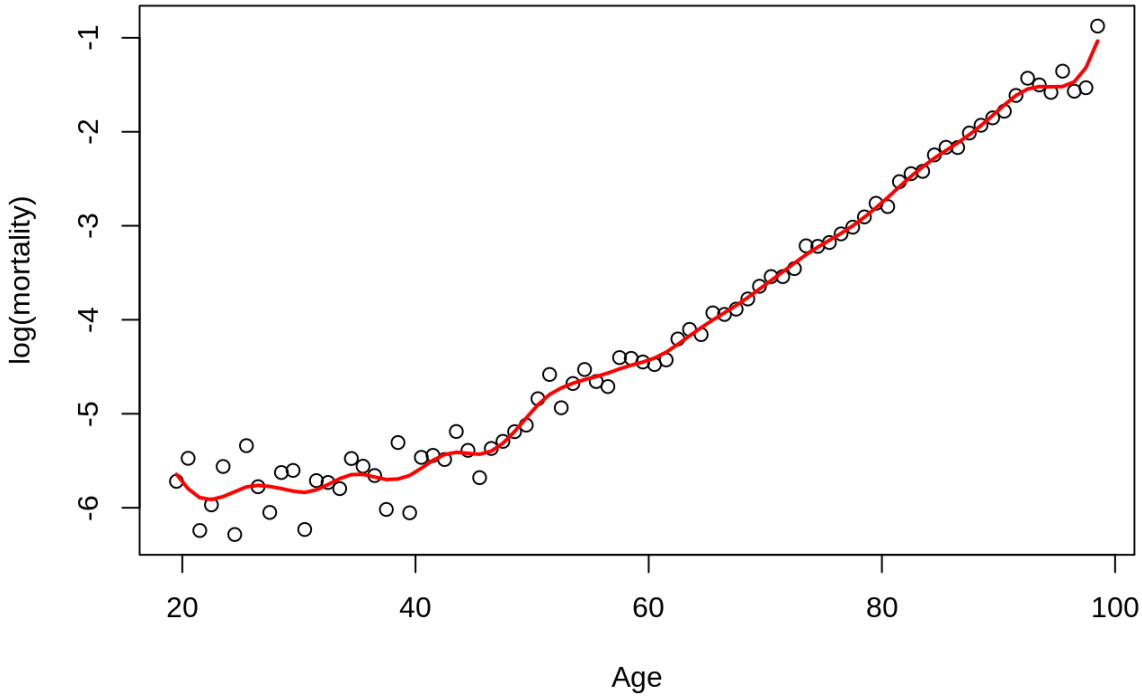
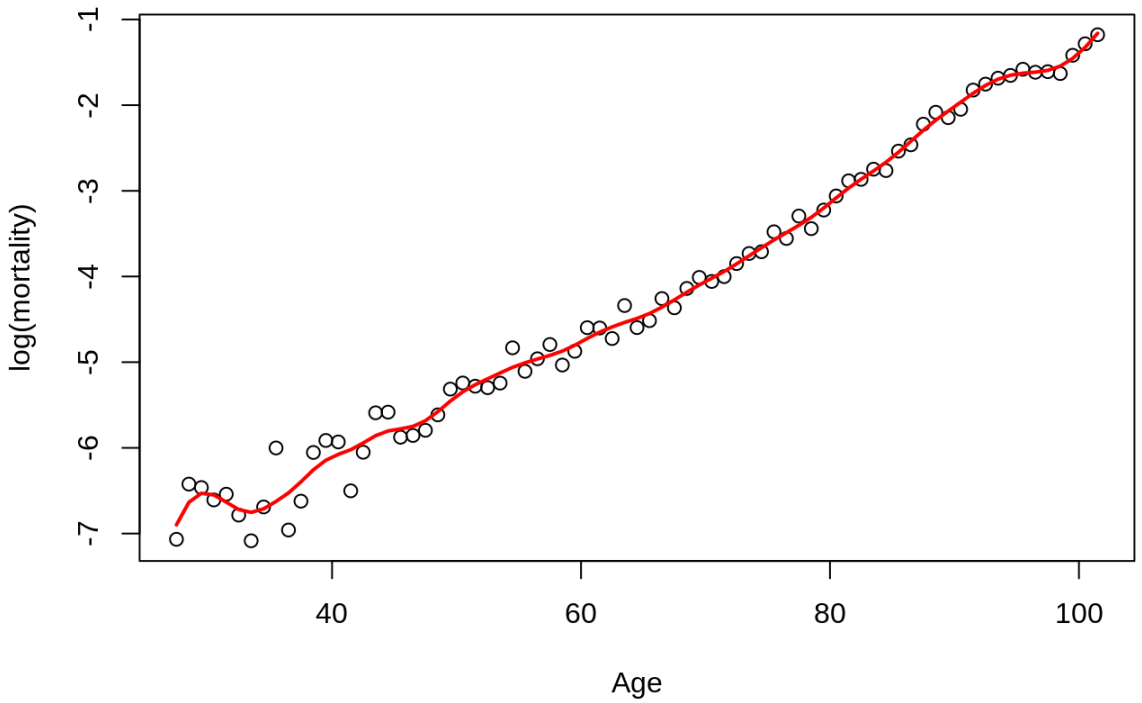


Abbildung 2: Referenztafel der Frauen aus den vorliegenden Datensätzen hergeleitet, logarithmierte Sterblichkeit pro Einzelalter



Wie zu erwarten ist, liegt die Sterblichkeit von Männern oberhalb der Sterblichkeit von Frauen bei einem Vergleich der Referenztafeln.

Mittels dieser beiden Referenztafeln wird im Folgenden die Anzahl der zu erwartenden Todesfälle für einen beliebigen Teilbestand berechnet. Diese A/E (Actual vs Expected)-Analyse gibt einen ersten, klassischen Eindruck von den Unter-

schieden in der Sterblichkeit der Teilbestände, ohne Korrelationen oder Bestandsverschiebungen zwischen den betrachteten Teilbeständen zu berücksichtigen.

Tabelle 4: Übersicht A/E nach dem Raucherstatus, altersabhängig für Frauen und Männer

Raucherstatus	Alters-band	beobachtete Todesfälle Frauen	beobachtete Todesfälle Männer	erwartete Todesfälle Frauen	erwartete Todesfälle Männer	A/E Frauen	A/E Männer
Nichtraucher	(27,37]	61	107	68.87	115.74	88.6%	92.5%
	(37,47]	135	146	160.21	182.93	84.3%	79.8%
	(47,57]	239	258	287.69	339.30	83.1%	76.0%
	(57,67]	379	414	464.78	551.88	81.5%	75.0%
	(67,77]	807	994	935.30	1220.04	86.3%	81.5%
	(77,87]	1935	2198	2038.47	2282.41	94.9%	96.3%
	(87,97]	1426	983	1422.49	992.05	100.2%	99.1%
Raucher	(27,37]	36	83	25.91	71.92	139.0%	115.4%
	(37,47]	83	141	51.37	94.94	161.6%	148.5%
	(47,57]	155	267	95.13	180.68	162.9%	147.8%
	(57,67]	198	361	117.64	222.82	168.3%	162.0%
	(67,77]	299	523	150.33	295.91	198.9%	176.7%
	(77,87]	253	341	153.15	249.66	165.2%	136.6%
	(87,97]	63	75	62.38	67.91	101.0%	110.4%

Man erkennt sehr deutlich den starken Einfluss des Rauchens auf die Sterblichkeit sowohl für Männer als auch Frauen und außerdem eine mit dem Alter ansteigende Übersterblichkeit bis zum Altersband der 68 bis 87-Jährigen

Vergleichbar zeigt folgende Tabelle die Bedeutung eines Alkoholmissbrauchs:

Tabelle 5: Einfluss eines Alkoholmissbrauchs auf die Sterblichkeit von Frauen und Männern

Alkoholmiss-brauch	beobachtete Todesfälle Frauen	beobachtete Todesfälle Männer	erwartete Todesfälle Frauen	erwartete Todesfälle Männer	A/E Frauen	A/E Männer
Nein	3125	3921	3463.04	4432.78	90.2%	88.5%
Ja	342	1617	196.08	1297.98	174.4%	124.6%

Bemerkenswerterweise ist die Übersterblichkeit von Frauen bei dem Vorliegen eines Alkoholmissbrauchs deutlich stärker ausgeprägt im Vergleich zu Männern.

Analog können die weiteren Variablen klassisch betrachtet werden, um einen ersten Eindruck von relevanten Kovariablen für die Sterblichkeit zu gewinnen

- A/E Analyse zur Referenztafel für verschiedene Segmente
    - Einkommen
    - BMI
    - Bildungsstatus
    - Familienstand etc.

7.2. Regression

Auf den obigen Datensatz, der für die actuarial-exposure-method oder auch Kalenderjahrmethode hergeleitet wurde, hat die Arbeitsgruppe bereits verschiedenste ML-Verfahren probeweise angewandt. Der Datensatz wird dabei mit folgenden Verfahren analysiert.

- Generalized Linear Models (GLM) mit und ohne Offsets
- Decision Trees (CART) mit und ohne Offsets
- Random Forests auf die Zellenergebnisse der Actuarial Exposure Method
- Gradient Boosting Machine auf die Zellenergebnisse A/E mit und ohne Offsets
- XGBoost mit genetischem Algorithmus zur Parameteroptimierung
- Survival-Cox-PH, Survival Trees

Dabei wurden in einer ersten Analyse folgende Variablen herangezogen: Alter, Geschlecht, Raucherstatus, Alkoholkonsum, Personenstand, Einkommen und Bildungsstatus.

Beispielhaft seien folgende Ergebnisse einzelner Verfahren aufgeführt:

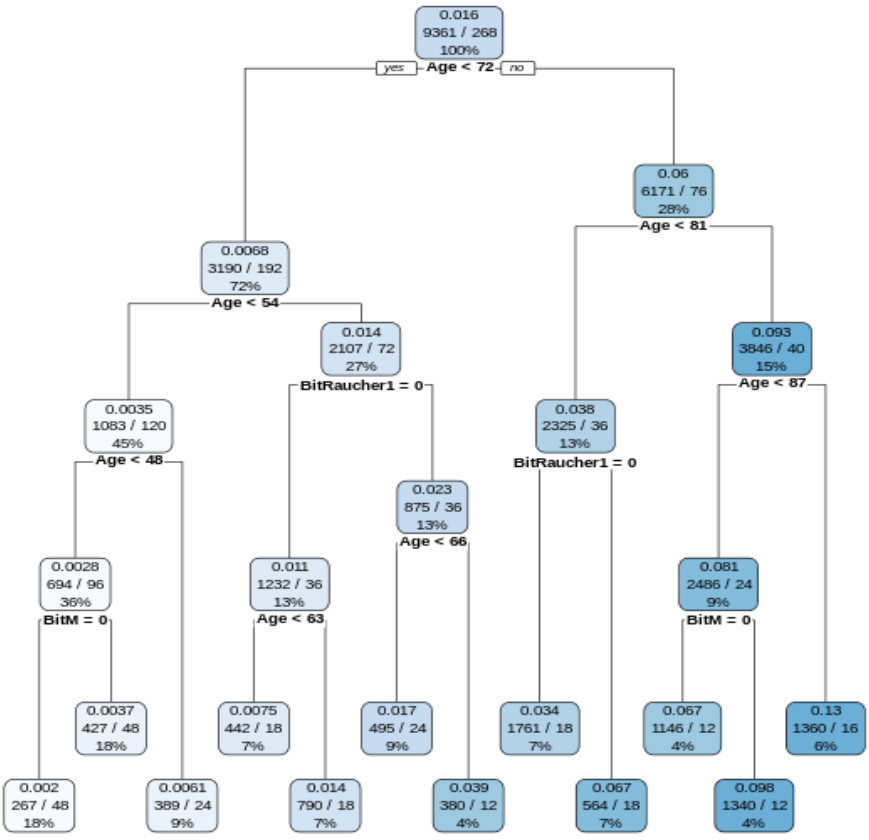


Abbildung 3: Ergebnisse eines CART-Modells mit den Variablen Alter, Geschlecht und Raucherstatus

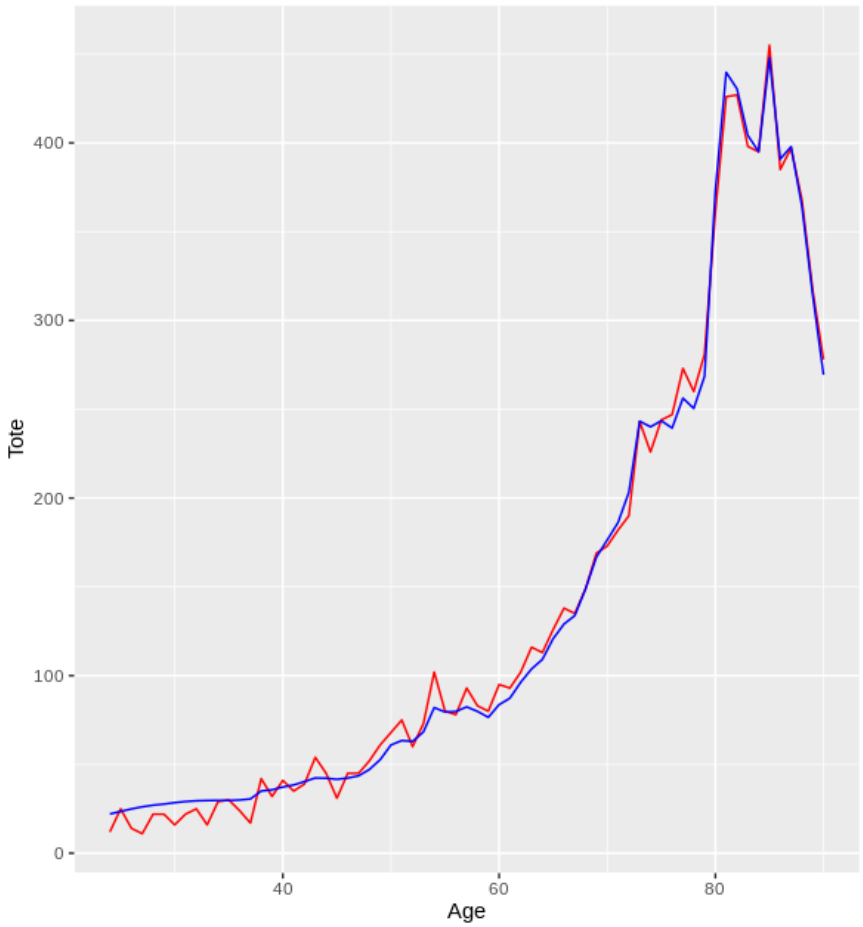


Abbildung 4: Vergleich der real beobachteten Todesfälle (rot) mit den aus dem Modell projizierten Anzahl von Todesfällen (blau)

Als Gütemaße wurden u. a. verwendet: Absoluter Fehler, relativer Fehler, Poisson Deviance und die bekannten Tuning-Verfahren.

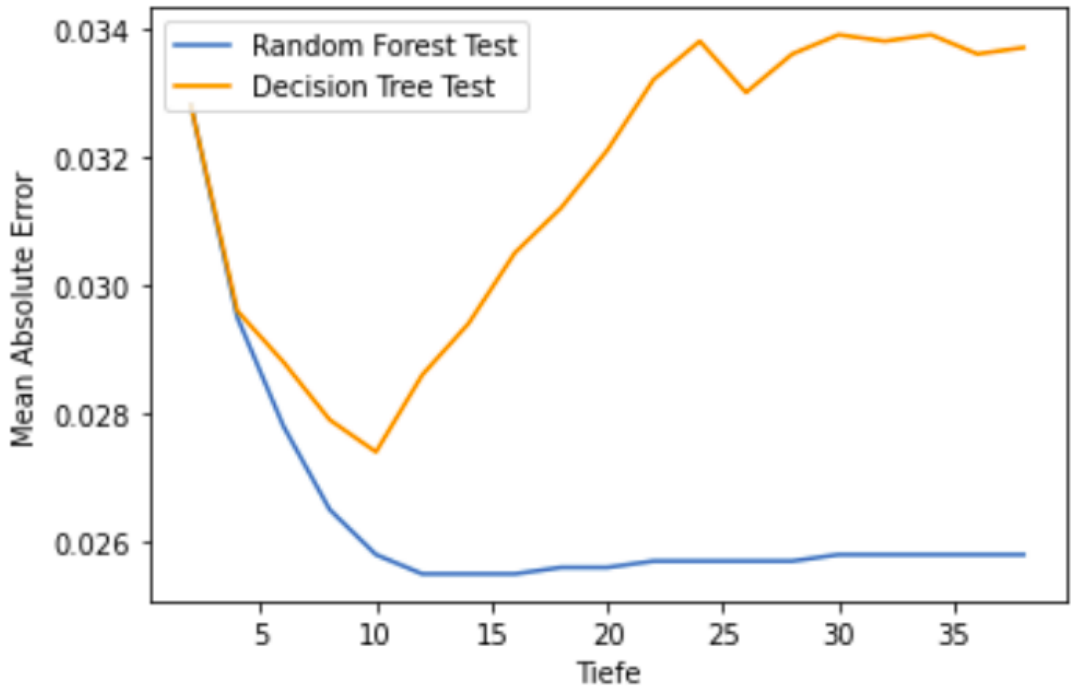


Abbildung 5: Hyperparameter-Tuning Random Forest im Vergleich zu einem Decision Tree

8. Publizierung von Ergebnissen

Dieser erste Teil des Ergebnisberichts mit der Beschreibung der Herleitung und Aufbereitung der NHANES-Daten wird klassisch seitens der DAV veröffentlicht. Die zugehörigen Skripte und Konfigurationsdateien finden sich auf dem github-Verzeichnis der DAV im Projektverzeichnis NHANES. Die Skripte sind unter verschiedensten Bedingungen getestet worden. Dennoch kann es zu Problemen und Fehlern in der Ausführung kommen. Die Arbeitsgruppenmitglieder stehen gerne für Fragen zur Verfügung.

9. Literatur, weitere Quellen und Links

[1] „Big Data in der Lebensversicherung", 2019, Hinweis des Ausschusses Lebensversicherung, Deutsche Aktuarvereinigung (DAV e. V.)

[2] National Center for Health Statistics, CDC (Center for Disease Control and Prevention, US): https://www.cdc.gov/nchs/

[3] NHANES-Datensatz: langjährige Datenerhebung der CDC in den USA, https://www.cdc.gov/nchs/nhanes/index.htm

[4] Linked Mortality File, NCHS, CDC, https://www.cdc.gov/nchs/data-linkage/mortality-public.htm

[5] Data-Use-Restrictions NHANES, CDC: https://www.cdc.gov/nchs/data_access/restrictions.htm

[6] Data-Use-Restrictions Linked Mortality File, CDC, siehe auch Beschreibung zum LMF: https://www.cdc.gov/nchs/data/datalinkage/public-use-2015-linked-mortality-file-description.pdf

[7] DGVFM-Datenbankprojekt: https://aktuar.de/forschung-und-transfer/datenbankprojekt/Seiten/default.aspx

[8] DAV-Arbeitsgruppe Actuarial Data Science, Anwendungsfall 3 Neuronale Netze treffen auf Mortalitätsprognose: https://aktuar.de/unsere-themen/big-data/anwendungsfaelle/Seiten/anwendungsfall3.aspx

[9] Ergebnisbericht des DAV-Ausschusses Actuarial Data Science (DAV e. V.), „Anwendung von künstlicher Intelligenz in der Versicherungswirtschaft", 2020, Köln