DAV
Deutsche
Aktuarvereinigung e.V.

Ergebnisbericht des Fachausschusses Altersversorgung

Data Science in der betrieblichen Altersversorgung

Köln, 21. Oktober 2024

Präambel

Der Pool *Data Science in der bAV* des Fachausschusses Altersversorgung der Deutschen Aktuarvereinigung e. V. (DAV) hat den vorliegenden Ergebnisbericht erstellt.[1]

Anwendungsbereich

Der Ergebnisbericht betrifft Aktuarinnen und Aktuare[2] die in den verschiedenen Durchführungswegen der bAV tätig sind.

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.[3]

Inhalt

Der Ergebnisbericht behandelt Fragestellungen zum Einsatz von Data Science in der betrieblichen Altersversorgung (bAV). Es werden zum einen grundsätzliche Aspekte zum Einsatz von Methoden von Data Science in der bAV erörtert. Dies betrifft insbesondere die rechtlichen Rahmenbedingungen sowie mögliche Datenquellen. Zudem wird eine Taxonomie der Daten erstellt. Zum anderen werden mögliche Anwendungsbeispiele in der bAV aufgezeigt.

Der Ergebnisbericht erhebt keinen Anspruch auf Vollständigkeit, sondern soll vielmehr einen Einstieg in die Thematik ermöglichen und an geeigneten Stellen auf weiterführende Quellen hinweisen.

Schlagworte

Ergebnisbericht, Altersversorgung, Actuarial Data Science

Verabschiedung

Dieser Ergebnisbericht ist durch den Fachausschuss Altersversorgung und den Ausschuss Actuarial Data Science am 21. Oktober 2024 verabschiedet worden.

---

[1] Der Ausschuss dankt dem Pool *Data Science in der bAV* ausdrücklich für die geleistete Arbeit, namentlich Dr. Sandra Blome, Dr. Sebastian Leipert, Clemens Sommer und Johannes Strenger.

[2] Auch wenn hier und im Folgenden die Aktuarinnen und Aktuare explizit genannt werden, spricht die DAV alle Geschlechter und Identitäten gleichermaßen an. Dies gilt auch für alle anderen hier genannten Personengruppen.

[3] 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.

This abstract summarises the report on findings „Data Science in der betrieblichen Altersversorgung“, which was approved by the DAV Pension committee on 21.10.2024.

Data science in occupational pension schemes

This report, prepared by the Data Science Working Group within the Pension Committee of the German Actuarial Association (DAV), explores the application of data science in occupational pension schemes. The study examines various aspects of integrating data science into occupational pension schemes, including the legal frameworks, potential data sources, and their classification. Key areas of focus include the use of machine learning and data analytics to address challenges such as the lack of pension coverage, the effects of data science on retirement timing, health, and labor performance. Additionally, the report outlines practical examples and proposes a taxonomy for data relevant to occupational pension schemes, involving primary, secondary, and supplementary data sources.

The report does not aim to be exhaustive but seeks to provide an introduction to the topic, facilitating further exploration and application of data science in the field of occupational pension schemes. It highlights the need for actuaries to balance technical innovation with compliance to regulations, such as the GDPR, while suggesting that historical data can be used for predictive analytics and the improvement of administrative processes. Future applications discussed include improving pension fund management, supporting employer and employee communication, and enhancing the precision of biometric bases of calculation.

Reports on findings are summaries of the results of work carried out by DAV committees or working groups,

- where their application can be freely decided upon within the framework of the code of conduct,
- that should inform discussion of the current opinion among actuaries or also among the broader public.

As working results of a single committee, they do not, for the time being, represent any recognised position within the DAV and do not comprise any actuarial standards of practice. In this respect they are clearly distinguishable from any standards of practice.

Inhaltsverzeichnis

1. Einleitung

1.1. Daten und betriebliche Altersversorgung

Daten sind der Treibstoff der Versicherungswirtschaft[4]. Dies gilt auch für die bAV, wobei hier noch zusätzliche Daten in Betracht gezogen werden müssen, die sich z.B. konkret aus dem Arbeitsverhältnis speisen.

Was kann man unter Data Science im Rahmen dieses Berichtes verstehen? Data Science wird als Schnittmenge folgender Punkte verstanden:

1. Domänenwissen
2. Datenmanagement
3. Datenanalyse

Unter dem **Domänenwissen** ist hierbei das Fachwissen rund um die bAV entscheidend. Neben dem (allgemeinen) aktuariellen Wissen ist hier detaillierteres Fachwissen erforderlich. Exemplarisch sei hierbei das Wissen genannt zu:

1.1. sozialversicherungs- und arbeitsrechtlichen Aspekten, wie z.B. der reinen Beitragszusage

1.2. handels- und steuerrechtlichen Vorgaben des Gesetzgebers, wie z.B. für die Bewertung von Versorgungszusagen

1.3. Wissen rund um Geschäftsvorfälle und deren Abwicklung (z.B. Liquidation) in der Praxis

Grundsätzlich betrifft das **Datenmanagement** Aktuare zum einen bei ihrer (Mit-) Arbeit an der (Weiter-)Entwicklung von Bestandsführungssystemen. Zum anderen können Daten aus weiteren Quellen und Systemen hinzukommen. Damit verbunden ist das Wissen um

2.1. die Datenintegration,

2.2. die Datenbereinigung,

2.3. Datenabzüge und

2.4. die Datenvisualisierung.

Der dritte Punkt, die **Datenanalyse**, umfasst im Wesentlichen die folgenden zwei Felder:

3.1. Data Analytics, also die Untersuchung von Datensätzen mit dem Ziel, geeignete Erkenntnisse daraus abzuleiten.

3.2. Maschinelles Lernen als ein Teilbereich der Künstlichen Intelligenz, in dem einem Computer die Fähigkeit zu lernen gegeben wird, ohne explizit programmiert zu werden.[5] Die lernenden Algorithmen gewinnen Erkenntnisse aus Daten und/oder unterstützen als Künstliche Intelligenz in Prozessen der bAV oder geben sogar Prognosen ab.

Betriebliche Altersversorgung umfasst dabei alle Arten der Absicherung von biometrischen Risiken, die durch das Betriebsrentengesetz (BetrAVG) als bAV definiert sind und betrifft damit alle Arbeitnehmer[6], die gemäß BetrAVG in den Genuss einer bAV kommen dürfen, sowie alle Leistungsempfänger einer bAV. Dadurch werden folglich auch alle Durchführungswege erfasst, alle

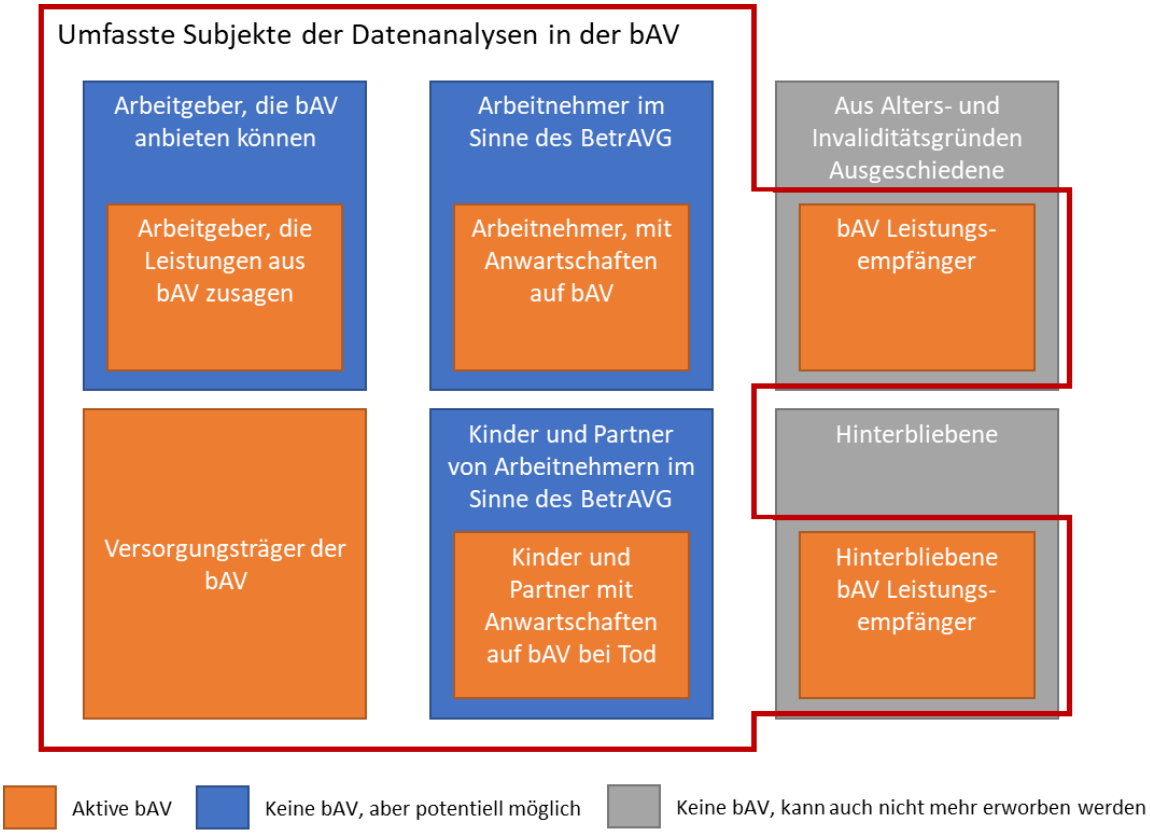
[4] Big Data Leben in der Lebensversicherung, Ergebnisbericht des Ausschusses Lebensversicherung, 19.09.2019

[5] Arthur Samuel (1959), S. https://link.springer.com/chapter/10.1007/978-1-4302-5990-9_1 abgerufen am 17.1.2024

[6] Einschließlich beherrschende Gesellschafter Geschäftsführer und Personen, die nicht Arbeitnehmer sind, denen aber nach § 17 Abs. (1) Satz 2 BetrAVG Leistungen der Alters-, Invaliditäts- oder Hinterbliebenenversorgung aus Anlass ihrer Tätigkeit für ein Unternehmen zugesagt worden ist.

Finanzierungsformen, alle Arten zur Gestaltung von Zusagen und Leistung. Damit umfasst werden auch die Daten aller durch eine bAV tangierten Subjekte wie Arbeitgeber, Arbeitnehmer, Hinterbliebene und Versorgungsträger[7] der bAV.

Das nachfolgende Bild veranschaulicht die datengetriebene Sicht auf die Subjekte für Datenanalysen zur bAV. Datengetrieben bedeutet in diesem Zusammenhang vor allem auch, dass historische Daten genutzt werden. So kann eine Person, die in der Vergangenheit in einem Beschäftigungsverhältnis gestanden hat, das grundsätzlich eine bAV nach dem BetrAVG erlaubt und die zum Zeitpunkt der Datenanalysen z.B. erwerbslos oder selbständig ist, Ziel dieser Datenanalysen sein. Dies gilt unabhängig davon, ob die Person tatsächlich eine bAV erworben hat.



Potenzielle Fragestellungen, denen die Data Science in der bAV nachgehen kann, sind u.a.:

- Gründe für das Fehlen einer bAV,
- Gründe der Höhe der Beiträge und der Leistung,
- Auswirkung einer bAV auf den Zeitpunkt des Leistungsbezugs,
- Auswirkung einer bAV auf die Gesundheit oder die Arbeitsleistung,
- Einfluss diverser Merkmale auf die biometrischen Rechnungsgrundlagen oder das Arbeitnehmerverhalten hinsichtlich Fluktuation oder Kapitalwahl,
- u.v.m.

[7] Der Begriff Versorgungsträger umfasst alle Einrichtungen, die eine bAV durchführen: Pensionskassen inkl. Zusatzversorgungskassen, Pensionsfonds, Unterstützungskassen, Lebensversicherungen, Arbeitgeber

Ergebnisse aus diesen Fragestellungen können Grundlagen bieten zu ökonomischen und praktischen Fragestellungen wie z.B.

- Prognosen über den Erfolg von bAV-Angeboten,
- Ermittlung potenzieller Zielgruppen für bAV-Angebote,
- Unterstützung in der Vertriebsansprache von Arbeitgebern und Arbeitnehmern durch gezielte und bedarfsgesteuerte Ansprachen,
- Unterstützung der Administration in der Kommunikation mit Arbeitnehmern,
- Bessere Schätzung der biometrischen Rechnungsgrundlagen,
- Prognosen zum Kapitalbedarf und der Risikobeurteilung von bAV-Beständen, sowie zur Validierung von Bewertungen,
- Bestandsverdichtungen, -analysen und -bereinigungen,
- Unterstützung bei der Migration von Altdaten in neue Zielsysteme,
- Gesundheitsvorsorge zur Prävention von Invalidität und Berufsunfähigkeit,
- u.v.m.

1.2. Vorgehen

Der vorliegende Ergebnisbericht wurde vom Pool zu Fragestellungen im Bereich Data Science in der bAV erstellt und trägt Informationen zur Nutzung von Data-Science-Methoden im Kontext der bAV zusammen.

Es wurde bei der Gründung des Pools keine konkrete Fragestellung an den Pool gerichtet, die es zu analysieren und ggf. zu lösen gilt. Die Aufgabe war mit einem hohen Freiheitsgrad versehen und bestand darin, festzustellen, ob es bereits Data-Science-Aktivitäten gibt, sowie mögliche Ansätze für Data Science in der bAV zu entwickeln.

Davon ausgehend hat sich der Pool auf drei Aspekte konzentriert:

- Evaluierung der rechtlichen Rahmenbedingungen für Data Science,
- Erfassen von möglichen Datenquellen,
- Entwicklung möglicher Anwendungen für Data Science.

Berücksichtigung fanden die Arbeiten anderer Arbeitsgruppen, um Doppelarbeiten zu vermeiden. An dieser Stelle weisen wir insbesondere auf die Arbeitsgruppe Bestandsmigration in der Lebensversicherung hin (gemeinsame Arbeitsgruppe der Ausschüsse Lebensversicherung und Actuarial Data Science) sowie auf den Ergebnisbericht des Ausschuss Lebensversicherung „Big Data in der Lebensversicherung“ vom 19.9.2019.

Im Kapitel 2 dieses Ergebnisbericht werden grundsätzliche Aspekte zum Einsatz von Data Science in der bAV erläutert. Dabei werden aktuelle und in Diskussion befindliche rechtliche Rahmenbedingungen erläutert. Im Kapitel 3 wird auf mögliche Datenquellen und eine Taxonomie eingehen. Im Kapitel 4 werden Anwendungsbeispiele vorgestellt und im abschließenden Kapitel 5 die wesentlichen Ergebnisse des Berichts zusammengefasst.

2. Rechtliche Rahmenbedingungen

Die rechtlichen Aspekte von Data Science in der Versicherung wurde von der Deutschen Aktuarvereiningung bereits in einigen Dokumenten diskutiert, die auch für die bAV von Interesse sind. Um doppelte Darstellungen zu vermeiden, wird im folgenden stichpunkthaft auf die relevanten Inhalte verwiesen:

- Ergebnisbericht des Ausschusses Actuarial Data Science: “Data Governance Act”, Köln, 18.10.2022 → Überblick und ausgewählte Inhalte über den Data Governance Act
- Ergebnisbericht des Ausschusses Actuarial Data Science: „Umgang mit Daten im Bereich Data Science“, Köln, 14. Februar 2020 → Hilfestellung im Umgang mit Daten ergänzend zu den Standesregeln für Aktuare (allgemeine Prinzipien und ethische Grundsätze)
- Ergebnisbericht des Ausschusses Actuarial Data Science: „Anwendung von Künstlicher Intelligenz in der Versicherungswirtschaft“, Köln, 14. Februar 2020 → Regulierung bzgl. Vertrauenswürdigkeit der eingesetzten KI-Systeme, Erklärbareit und Interpretierbarkeit verwendeter Algorithmen

2.1. Aktuelle rechtliche Rahmenbedingungen

Die im Jahr 2018 auf europäischer Ebene erlassene Datenschutzgrundverordnung (DSGVO) gilt als Verordnung unmittelbar auch in Deutschland. Ziel der DSGVO ist der Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten und die Schaffung eines gesetzlichen Rahmens zum freien Verkehr solcher Daten.[8] Der Anwendungsbereich umfasst im Wesentlichen alle Verarbeitungen personenbezogener Daten durch eine elektronische Datenverarbeitung sowie strukturierte analoge Datensammlungen im beruflichen oder wirtschaftlichen Kontext. Somit ist sie für den Bereich Data Science in jedem Fall zu beachten.

Personenbezogene Daten sind alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen. Beispiele sind Name, Anschrift und E-Mail-Adresse. Im Fall eines Arbeitsverhältnisses sind weitere Beispiele die Personalnummer, das Gehalt, die Bankverbindung sowie der berufliche Werdegang.[9]

Personenbezogene Daten, die anonymisiert, verschlüsselt oder pseudonymisiert wurden, aber zur erneuten Identifizierung einer Person genutzt werden können, bleiben personenbezogene Daten und fallen in den Anwendungsbereich der DSGVO.[10] Personenbezogene Daten, die in einer Weise anonymisiert worden sind, dass die betroffene Person nicht oder nicht mehr identifiziert werden kann, gelten nicht mehr als personenbezogene Daten. Damit die Daten wirklich anonymisiert sind, muss die Anonymisierung unumkehrbar sein.

Gemäß Artikel 5 der DSGVO gilt die Zweckbindung. Personenbezogene Daten müssen für festgelegte, eindeutige und legitime Zwecke erhoben werden und dürfen nicht in einer mit diesem Zwecken nicht zu vereinbarenden Weise weiterverarbeitet werden. Eine Weiterverarbeitung für wissenschaftliche und statistische Zwecke gilt nicht als unvereinbar mit den ursprünglichen Zwecken.[11]

Die DSGVO legt fest, dass personenbezogene Daten nur dann verarbeitet werden dürfen, wenn dies durch eine bestimmte Rechtsgrundlage oder eine Einwilligung der Person erlaubt ist. Die Rechtsgrundlage im Fall eines Arbeitsverhältnisses findet sich im Bundesdatenschutzgesetz

[8] Artikel 1 Absatz 1 DSGVO

[9] S. https://www.lexoffice.de/lohn/wissen/datenschutz-dsgvo-mitarbeiter/ abgerufen am 12.7.2023

[10] S. https://commission.europa.eu/law/law-topic/data-protection/reform/what-personal-data_de abgerufen am 12.7.2023

[11] Art. 5 Abs. (1) Buchstabe b) DSGVO https://dsgvo-gesetz.de/ abgerufen am 22.07.2024

(BDSG). Das BDSG bestimmt, dass Arbeitgeber auch ohne Einwilligung der Mitarbeiter solche personenbezogenen Daten verarbeiten dürfen, die für die Aufnahme, Durchführung oder Beendigung eines Arbeitsverhältnisses erforderlich sind.[12]

Die DSGVO setzt hohe Hürden für den Einsatz von Data Science in der bAV. Obwohl eine Weiterverarbeitung von Daten für wissenschaftliche und statistische Zwecke möglich ist, wird es in der Praxis oft daran scheitern, dass Daten aus den unterschiedlichen Quellen nicht direkt zusammengeführt werden dürfen und Dritten zum Zwecke einer Analyse nicht bereitgestellt werden dürfen. Die Analyse der eigenen Daten von Arbeitgebern und Versorgungsträgern ist jedoch möglich, beispielsweise wenn diese für die Durchführung des Arbeitsverhältnisses oder der bAV erforderlich sind. So ist es denkbar, dass Data Analytics in einer Pensionskasse eingesetzt wird, um beispielsweise die Verarbeitungsprozesse zu optimieren und damit die Verwaltungskosten zu senken. Externe Daten können den Analysen nur in aggregierter Form zugespielt werden.

Im Rahmen einer Auftragsdatenverarbeitung ist es möglich, dass die Verarbeitung der Daten zur Durchführung von Arbeitsverhältnissen durch Dritte übernommen werden kann. Diese Dienstleister verfügen damit über Daten zu Arbeitsverhältnissen bei verschiedenen Arbeitgebern, dürfen diese jedoch nur im gesetzten Rahmen des Auftrags verarbeiten und insbesondere auch die Daten der verschiedenen Arbeitgeber nicht zusammenführen oder analysieren. Dies gilt insbesondere für die vom Pool bAV in Abschnitt 3.2 betrachteten bAV-Administratoren.

2.2. Aktueller Stand der weiteren europäischen Gesetzgebung

Die EU-Kommission arbeitet seit einigen Jahren intensiv an übergreifenden Regelungen zur Nutzung von Verfahren zur Künstlichen Intelligenz. Diese Regelungen werden flankiert von diversen weiteren umfassenden Regelungen zur Nutzung von Daten. Die Regelungen finden sich fast immer wieder in einer Umsetzung der Regelungen und Empfehlungen in nationalem Recht. Die Versicherungsbranche ist von diesen Regelungen ebenfalls betroffen. Sie haben damit für die Versicherungsbranche eine praktische Umsetzungsrelevanz über die reine Nutzung von Künstlicher Intelligenz und Machine Learning-Verfahren hinaus.

An dieser Stelle werden die derzeit wichtigsten Verfahren und Regelungen im Überblick genannt und die Fundstelle bei der Kommission aufgezeigt. Alle wesentlichen Verfahren wurden während der Arbeit des Pools fertiggestellt und im Ergebnisbericht entsprechend mit Stand Juli 2024 berücksichtigt. Jedoch sollte bedacht werden, dass die Dynamik der Gesetzgebung zu diesen Regelungen sehr hoch ist und die hier dargestellten Ergebnisse auch nach kurzer Zeit einer Validierung bedürfen.

2.2.1. Ethics Guidelines for Trustworthy AI

Bei den *Ethics Guidelines for Trustworthy AI*[13] (zu Deutsch: Ethik-Leitlinien für eine vertrauenswürdige KI) handelt es sich um eine Leitlinie, deren „Ziel (…) die Förderung einer vertrauenswürdigen KI“ ist (siehe Seite 2 Abs. (1) der Leitlinie)[14]. Die Leitlinien beschreiben, welche Komponenten während des gesamten Lebenszyklus des Systems erfüllt sein sollen. Sie legt außerdem einen „Rahmen für die Verwirklichung einer vertrauenswürdigen KI“ fest, beschreibt Prinzipien und Grundsätze, wie z.B. die Achtung der menschlichen Autonomie, Schadensverhütung, Fairness und Erklärbarkeit, die zu befolgen sind und versteht sich als „Hilfestellung für die mögliche Umsetzung dieser Prinzipien in soziotechnischen Systemen“.

[12] S. https://www.lexoffice.de/lohn/wissen/datenschutz-dsgvo-mitarbeiter/ abgerufen am 12.7.2023

[13] Allgemeine Informationen zum Ethikleitfaden
https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai abgerufen am 24.07.2023

[14] Ethikleitfaden in der Fassung vom 08.11.2019
https://op.europa.eu/de/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1 abgerufen am 24.07.2023

Erstellt von der High-Level Expert Group on Artificial Intelligence (HLEG), European Commission, Brüssel, sind die *Ethics Guidelines for Trustworthy AI* keine Rahmenrichtlinie oder Grundverordnung und damit nicht rechtsverbindlich. Ziel der Leitlinie ist es, Prinzipien und Anforderungen an KI-Systeme zu beschreiben. Die Gruppe identifiziert *drei Komponenten*, die ein KI-System erfüllen muss[15]:

1) Sie soll *rechtmäßig* sein und alle geltenden Rechts- und Verwaltungsvorschriften einhalten;
2) sie soll *ethisch* sein und ethischer Grundsätze und Werte achten;
3) sie soll *robust* sein, sowohl aus technischer Sicht als auch unter Berücksichtigung des sozialen Umfelds, da KI-Systeme selbst in guter Absicht unbewusst Schaden verursachen können.

Hinsichtlich der Ethik sollen sich KI-Systeme an *vier ethischen Grundätzen* (ethische Imperative)[16] orientieren. Diese sind

1) Achtung der menschlichen Autonomie;
2) Schadenverhütung;
3) Fairness;
4) Erklärbarkeit.

Insgesamt ergibt sich durch den Leitfaden zwar keine rechtlich bindende Vorschrift für Data Science in der bAV, allerdings empfiehlt es sich bei der „Entwicklung, Einführung und Nutzung von KI-Systemen“ die im Zusammenhang mit der Einhaltung der vier ethischen Grundsätze formulierten *Anforderungen an vertrauenswürdige KI* (zu) erfüllen[17]:

1) Vorrang menschlichen Handelns und menschlicher Aufsicht;
2) technische Robustheit und Sicherheit;
3) Schutz der Privatsphäre und Datenqualitätsmanagement;
4) Transparenz;
5) Vielfalt, Nichtdiskriminierung und Fairness;
6) gesellschaftliches und ökologisches Wohlergehen;
7) Rechenschaftspflicht.

2.2.2. Artificial Intelligence Act

Der *Artificial Intelligence Act* (AI Act, zu Deutsch: Gesetz über künstliche Intelligenz oder kurz AI-Gesetz)[18,19,20] ist eine Regulierung der Europäischen Union, die am 21.04.2021 von der Europäische Kommission vorgeschlagen und nach 3 jähriger Verhandlung durch den Europäischen Rat

[15] Ethik Leitlinien für eine Vertrauenswürdige KI, RZ (15) https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai abgerufen am 07.07.2024

[16] Ethik Leitlinien für eine Vertrauenswürdige KI, RZ (48) https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai abgerufen am 07.07.2024

[17] Ethik Leitlinien für eine Vertrauenswürdige KI, RZ (58) https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai abgerufen am 07.07.2024

[18] Zusammenfassung des AI-Gesetzes https://artificialintelligenceact.eu/de/high-level-summary/ abgerufen am 26.05.2024

[19] Das vollständige Gesetz in der Fassung vom 19.04.2024 kann hier heruntergeladen werden: https://artificialintelligenceact.eu/de/das-gesetz abgerufen am 26.05.2024

[20] EU AI Act https://artificialintelligenceact.eu/de/ai-act-explorer/ abgerufen am 07.07.2024

am 21.05.2024 angenommen wurde. Beim AI Act handelt es sich um eine schwergewichtige Regulierung der Nutzung von Künstlicher Intelligenz im Europäischen Raum. Er regelt das Inverkehrbringen, die Inbetriebnahme und Verwendung von KI in der EU durch Gruppierung von KI-Anwendungen nach potenziellem Risiko. Der AI Act verfolgt unter anderem das Ziel, Diskriminierungen auf Grund des Einsatzes von Künstlicher Intelligenz zu vermeiden. Diese Verordnung wird nach ihrer Verabschiedung in europäisches Recht umzusetzen sein. Für Versicherer sind die darin enthaltenen Vorschriften zu berücksichtigen, zumal empfindliche Strafen drohen (u.a. Artikel 99 der Verordnung[21]), sofern KI-Systeme eingesetzt werden, die z.B. das Verbot der in Artikel 5 genannten Verfahren missachten. Insoweit ist der AI Act auch für Data Science in der bAV relevant.

Die genaue Wirkung auf die Versicherungsbranche wird sich zum Teil erst in der Zukunft abschätzen lassen. Wichtig ist hierbei u.a., welche Verfahren und Prozesse als Bestandteil eines Hochrisikosystems gelten werden. Als Hochrisiko-KI-Systeme werden im Art. 6, Abs. 2 die Systeme klassifiziert, die im Anhang III der Verordnung gelistet sind. Im Rahmen des Gesetzgebungsverfahren hat das Parlament den Kommissionsvorschlag deutlich angepasst und erweitert und damit das Geschäftsfeld der Versicherer in den Bereich der Hochrisiko-KI-Systeme aufgenommen: Mit dem Einschub, „KI-Systeme, die bestimmungsgemäß für die Risikobewertung und Preisbildung in Bezug auf natürliche Personen im Fall von Lebens- und Krankenversicherungen verwendet werden sollen“ zählen diese gemäß Anhang III Nr. 5 C zu den hochriskanten AI-Systemen.[22]

Da es Überschneidungen in den Tätigkeiten von Lebensversicherungen und Versorgungsträgern der bAV gibt, wird diese Einstufung auch im Umfeld der bAV Auswirkungen haben.

Unabhängig von den Verpflichtungen, die sich aus einer Klassifizierung als Hochrisikosystem ergeben, sind generell gemäß Artikel 50 des AI Act Transparenzpflichten zu erfüllen. Dies gilt insbesondere für die Fälle, in denen KI-Systeme zum Einsatz kommen, in denen natürliche Personen mit diesen Systemen interagieren. Denkbar wäre dies z.B. bei Portalen für die bAV. In diesen Fällen muss der Nutzende darüber informiert werden, dass eine Interaktion mit einem KI-System stattfindet. [23]

2.2.3. Data Governance Act

Die EU hat erkannt, dass das gesellschaftliche und wirtschaftliche Potenzial der Daten enorm ist, aber das Potential nicht ausreichend genutzt wird. Sie identifiziert eine Reihe von Hindernissen für eine Verbesserung des Datenaustauschs. Dazu zählen laut EU „geringes Vertrauen in den Datenaustausch, Probleme im Zusammenhang mit der Weiterverwendung von Daten des öffentlichen Sektors und Datenerhebung für das Gemeinwohl sowie technische Hindernisse“.[24]

Der Data Governance Act[25] (DGA) setzt es sich zum Ziel, mehr Daten zur Verfügung zu stellen. Dazu wird die Weiterverwendung von öffentlich gespeicherten und geschützten Daten geregelt und der Datenaustausch durch Regulierung von Datenintermediären gefördert. Zusätzlich wird der Datenaltruismus geregelt. Zitat: „Bei Data Altruism geht es um Einzelpersonen und Unternehmen, die ihre Zustimmung oder Erlaubnis erteilen, Daten zur Verfügung zu stellen, die sie – freiwillig und ohne Belohnung – erzeugen, um im öffentlichen Interesse verwendet zu werden.“

---

[21] Artikel 99 EU Artificial Intelligence Act https://artificialintelligenceact.eu/de/article/99/ abgerufen am 07.07.2024

[22] Anhang III Nr. 5 c) EU Artificial Intelligence Act https://artificialintelligenceact.eu/de/article/99/ abgerufen am 07.07.2024

[23] Artikel 50 EU Artificial Intelligence Act https://artificialintelligenceact.eu/de/article/99/ abgerufen am 07.07.2024

[24] Erläuterungen zum Data Governance Act https://digital-strategy.ec.europa.eu/de/policies/data-governance-act-explained abgerufen am 28.01.2024

[25] Data Governance Act in der Fassung vom 23.06.2022 https://digital-strategy.ec.europa.eu/de/policies/data-governance-act abgerufen am 25.07.2023

Die Bereitstellung der Daten darf nur durch öffentliche Stellen erfolgen, die technisch so ausgestattet sind, dass Datenschutz, Privatsphäre und Vertraulichkeit gewahrt bleiben.

Der im Data Governance Act definierte Datenaltruismus beinhaltet insbesondere, dass Daten allgemein zugänglich gemacht werden. Dies flankiert der DGA im Artikel 4 mit einem grundsätzlichen Verbot von Ausschließlichkeitsvereinbarungen, von dem nur sehr eingeschränkt abgewichen werden kann.

In Bezug auf Data Science für die bAV können unter anderem die folgenden öffentlichen Stellen Datenquellen besitzen, die von Interesse für die Forschung an Daten zur bAV sind:

- BaFin und andere Aufsichtsbehörden
- Deutsche Rentenversicherung
- Zentrale Zulagenstelle für Altersvermögen (ZfA)
- Zentrale Stelle für Digitale Rentenübersicht (ZfDR)

Da der Data Governance Act einen Rechtsrahmen setzt, der es öffentlichen Stellen erlaubt, Daten für die Forschung bereitzustellen, ist er für Data Science in der bAV relevant.

Die Data Governance trat am 23. Juni 2022 in Kraft und gilt nach einer Nachfrist von 15 Monaten seit September 2023.

2.2.4. Digital Services Act

Die Europäische Kommission hat einen Vorschlag für eine Verordnung[26] vorgelegt, um eine „Harmonisierung der Bedingungen für die Entwicklung innovativer grenzüberschreitender digitaler Dienste in der Union bei gleichzeitiger Wahrung eines sicheren Online-Umfelds (…) auf Unionsebene“[27] zu erreichen. Ziel ist es, ein einheitliches Schutzniveau für die EU-Bürger zu erreichen. Hervorzuheben ist insbesondere ein besserer Schutz der Grundrechte[28] sowie eine größere Auswahl an Waren und Dienstleistungen durch die Schaffung eines einheitlichen grenzüberschreitenden Rechtsrahmens.[29]

Für Anbieter digitaler Dienste bedeutet dies zum einen höherer Aufwand hinsichtlich des Schutzes der Grundrechte, strengere Aufsicht durch Behörden und den Aufbau weiterer Maßnahmen z.B. zur Vermeidung und Löschung illegaler Daten. Gleichzeitig wird durch die EU-weite Harmonisierung der Vorschriften darauf abgezielt, die Rechtssicherheit für die Anbieter zu erhöhen.

Hervorzuheben ist der Artikel 31 (Datenzugang und Kontrolle) bei dem sehr große Online-Plattformen der Forschung Zugriff auf die Kerndaten erlauben müssen, um das Fortschreiten von Online-Risiken nachvollziehen zu können.

Im Übrigen wird durch die Harmonisierung ein einheitlicher Rechtsrahmen insbesondere auch für gewerbliche Nutzer geschaffen, um Zugang zu einer größeren Auswahl an Dienstleistungen sowie EU-weiten Märkten über Plattformen zu erhalten.

---

[26] Gesetz über digitale Dienste (Entwurf) in der Fassung vom 15.12.2020 https://eur-lex.europa.eu/legal-content/de/TXT/?qid=1608117147218&uri=COM%3A2020%3A825%3AFIN abgerufen am 25.07.2023

[27] Siehe Nr.2 Absatz Subsidiarität der Begründung zum Gesetz über digitale Dienste (Entwurf) in der Fassung vom 15.12.2020 https://eur-lex.europa.eu/legal-content/de/TXT/?qid=1608117147218&uri=COM%3A2020%3A825%3AFIN abgerufen am 25.07.2023

[28] Charta der Grundrechte der EU in der Fassung vom 26.10.2012 https://ec.europa.eu/info/aid-development-cooperation-fundamental-rights/your-rights-eu/eu-charter-fundamental-rights_de abgerufen am 25.07.2023

[29] Erläuterungen zum Gesetzesvorschlag https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_de abgerufen am 25.07.2023

Das Gesetz ist am 16.11.2022 verabschiedet worden.[30] Allgemeiner Geltungsbeginn des Gesetzes über digitale Dienste ist der 17.02.2024.

Nach Einschätzung des Pools ist das Gesetz zwar grundsätzlich relevant für die Anbieter digitaler Dienste im Kontext der bAV, jedoch nicht unmittelbar relevant für Data Science in der bAV.

2.2.5. Digital Markets Act

Der „Digital Markets Act" trat am 01.11.2022 in Kraft[31] und wird seit dem 02.05.2023 angewendet[32]. Die Europäische Kommission schaffte damit eine Verordnung zum offenen und fairen Umgang auf digitalen Märkten. Adressaten des „Gesetzes über digitale Märkte"[33] sind große Plattformen, die auf Grund ihrer Marktmacht als Zugangstor bzw. Gatekeeper (Originalwortlaut aus der Begründung des Gesetzes) zwischen gewerblichen Nutzern und Endnutzern fungieren. Auf Grund ihrer Größe und des daraus resultierenden Netzwerkeffekts sowie auf Grund des umfassenden Trackings und Profilings von Endnutzern sind Gatekeeper in der Lage, Marktzutrittsschranken zu verstärken (siehe Kapitel 1, Begründung zum Vorschlag der Verordnung).

Das Gesetz definiert, ab wann eine Plattform in den Geltungsbereich des Gesetzes fällt (Artikel 3), und schreibt verbindlich vor, welche Verbote und Gebote für Gatekeeper gelten (Artikel 5 und 6). Dazu zählt beispielsweise nach Artikel 5 (3) die Regelungen, die es den gewerblichen Nutzern der Plattform des Gatekeepers ermöglicht, Verträge mit ihren Kunden auch außerhalb der Gatekeeper-Plattform abzuschließen. Unter anderem ist auch geregelt, dass es nach Artikel 5 (2) a) nicht zulässig ist, personenbezogene Daten aus unterschiedlichen Diensten des Gatekeepers oder aus Daten von Diensten Dritter zusammenzuführen.

Der Digital Markets Act ist vermutlich nicht relevant für Data Science in der bAV, sollte aber beobachtet werden, da sich im Europäischen Raum auch spezialisierte Plattformdienste etablieren. So fallen z.B. die Intermediäre, die im Rahmen der Digitalen Rentenübersicht Meldungen für die angeschlossenen Unternehmen an die ZfDR übernehmen unter Plattformdienste.

2.2.6. Data Act

Der Data Act[34] ist Bestandteil der europäischen Datenstrategie und wird von der europäischen Kommission vor den Hintergrund „eines sektorübergreifenden Governance-Rahmens für den Datenzugang und die Datennutzung"[35] geschaffen. Es soll ein Regelungsrahmen geschaffen werden, der innovationsfreundlich ist, eine bessere Datenübertragbarkeit gewährleistet und einen fairen Zugang zu Daten ermöglicht. Damit wird ein Rechtsrahmen geschaffen, der „das Potenzial von Daten und digitalen Technologien einschließlich künstlicher Intelligenz zum Vorteil der Gesellschaft und der Wirtschaft besser nutzt".[36]

---

[30] Gesetz über digitale Dienste https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:32022R1925 abgerufen am 24.01.2024

[31] Erläuterungen zum Gesetzesvorschlag https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2349 abgerufen am 22.01.2024

[32] About the Digital Markets Act: https://digital-markets-act.ec.europa.eu/about-dma_en abgerufen am 25.07.2023

[33] Gesetz über digitale Märkte (Entwurf) in der Fassung vom 15.12.2020 https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:32022R1925 abgerufen am 28.01.2024

[34] Data Act in der Fassung endgültigen Fassung vom 22.12.2023 https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:32023R2854 abgerufen am 28.01.2024

[35] Siehe Nr. 1 Absatz 3 der Begründung zum Data Act vom 23.02.2022 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2022%3A68%3AFIN abgerufen am 28.01.2024

[36] Weiterführende Informationen https://digital-strategy.ec.europa.eu/de/policies/data-act abgerufen am 28.01.2024

Bestandteil des Data Acts ist unter anderem die Erleichterung des Datenzugangs und der Datennutzung für Verbraucher und Unternehmen. Das beinhaltet insbesondere auch die bei Nutzung von Produkten oder den damit verbundenen Diensten erzeugten Daten.

Der Fokus des Datenaustauschs liegt in der Nutzung von Daten aus Geräten. Der Data Act ist daher vermutlich nicht unmittelbar für Data Science in der bAV relevant (vgl. hierzu auch die Ergebnisse des Fachausschuss Actuarial Data Science mit Sitzung am 05.07.2022).

Das Gesetz wurde durch den Rat der Europäischen Union am 27.11.2023 verabschiedet, trat am 11.01.2024 in Kraft und ist anzuwenden ab dem 12.09.2025.[37]

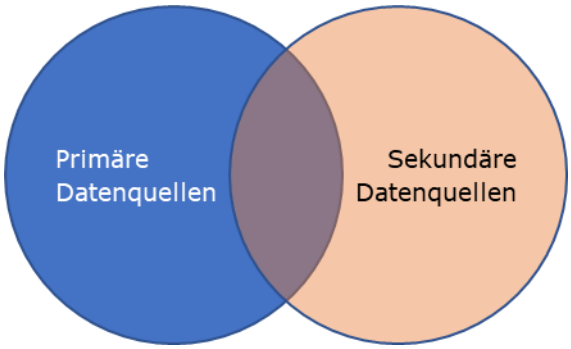
[37] https://www.europarl.europa.eu/news/de/press-room/20230310IPR77226/datengesetz-neue-regeln-fur-fairen-zugang-zu-und-nutzung-von-industriedaten abgerufen am 25.07.2023

3. Daten

3.1. Einführung

In dem folgenden Abschnitt bezeichne der Begriff **primäre Datenquellen** solche, bei denen die Daten sich unmittelbar Versorgungszusagen sowie Arbeitsverhältnissen, in denen die Zusage erteilt wurde, zuordnen lassen. Diese Daten sind u.a. Daten zur Person des Versorgungsberechtigten, zu ihrem Beschäftigungsverhältnis und zu ihrer Versorgung. Primäre Datenquellen finden sich typischerweise bei Versorgungsträgern wie z.B. Pensionskassen, Pensionsfonds, Lebensversicherern oder Unterstützungskassen sowie beim Arbeitgeber. Arbeitgebern kommt hier eine besondere Rolle zu, da sie im Zusammenhang mit der Erteilung von unmittelbaren Zusagen (Direktzusagen) auch die Rolle des Versorgungsträgers einnehmen. Daher unterscheiden wir im Folgenden zwischen externe Versorgungsträgern sowie Arbeitgebern mit Direktzusagen und Arbeitgebern ohne Direktzusagen[38].

Sekundäre Datenquellen seien dann solche, die über einen mittelbaren Bezug zu Arbeitsverhältnissen oder Versorgungszusagen verfügen. Dies können z.B. Daten des Erwerbslebens sein oder Daten zum Zeitpunkt des Todes. Sekundäre Daten und primäre Daten verfügen über eine nicht leere Schnittmenge. Als Beispiel kann der Verlauf eines Arbeitsverhältnisses dienen. Daten zum Verlauf des Arbeitsverhältnisses sind Bestandteil von primären Datenquellen. Sekundäre Datenquellen, wie z.B. von Sozialversicherungsträgern, verfügen ebenfalls über Verläufe von Arbeitsverhältnissen allerdings ohne Kenntnis darüber, ob gleichzeitig eine bAV besteht. Es ist zu prüfen, ob sich diese Datenquellen eignen, um z.B. Erkenntnisse aus primären Quellen zu extrapolieren. Typische Quellen für sekundäre Daten sind Daten von Behörden, z.B. dem Arbeitsamt oder der Deutschen Rentenversicherung und von Krankenversicherern.



Mit **anreichernden Datenquellen** bezeichnen wir Datenquellen, die nicht in Bezug zu Arbeitsverhältnissen oder Versorgungszusagen (primäre Daten) stehen. Hier bleibt konkret zu prüfen, ob diese Datenquellen genutzt werden. Als Beispiel diene hier der NHANES Datensatz[39].

Erkenntnisse aus Daten ergeben sich in der Regel durch Analyse konkreter Daten. Dem Pool stand leider keine primäre Datenquelle zur Verfügung. Daher hat die Arbeitsgruppe zunächst einen ersten Entwurf einer **Taxonomie** entwickelt. Dieses Vorgehen ist als hypothesenbasierter Ansatz grundsätzlich schwerer, als Erkenntnisse aus konkreten Daten zu entwickeln und birgt das Risiko, dass „an der Realität vorbei“ Aussagen getroffen werden.

[38] Sofern Arbeitgeber gleichzeitig die Aufgaben eines externen Versorgungsträgers (z.B. einer pauschaldotierten Unterstützungskasse) übernehmen, sind die vom Arbeitgeber verwalteten Daten zwei Datenquellen zuzuordnen: den Daten des Arbeitgebers und den Daten des externen Versorgungsträgers

[39] NHANES-Datensatz: langjährige Datenerhebung der CDC (Center for Disease Control and Preven-tion, US) in den USA, https://www.cdc.gov/nchs/nhanes/index.htm und auch Abschnitt **Fehler! Verweisquelle konnte nicht gefunden werden.**

Im Abschnitt 3.2 dieses Kapitels geht der Ergebnisbericht auf die verschiedenen primären Datenquellen und in Abschnitt 3.3 auf sekundäre Datenquellen ein. Im Abschnitt 3.4 wird eine Taxonomie primärer Datenquellen vorgestellt. Daraufhin werden im Abschnitt 3.5 anreichernde Datenquellen behandelt.

3.2. Primäre Datenquellen

Die folgenden Einrichtungen verfügen über primäre Datenquellen zur bAV:

1. Arbeitgeber
2. Versorgungsträger
3. bAV-Administratoren
4. Hybride Einrichtungen der bAV

Dabei sind bAV-Administratoren Einrichtungen, die für Arbeitgeber die Verwaltung der bAV übernehmen und dabei als eine Aufgabe die Funktion des Intermediär zwischen Arbeitgeber, Arbeitnehmer und Versorgungsträgern wahrnehmen. Unter bAV-Administratoren sollen in diesem Zusammenhang auch Einrichtungen fallen, die Bilanzgutachten für Unternehmen erstellen. Hybride Einrichtungen sind solche, die als Versorgungsträger zusätzlich zur Verwaltung des Versorgungsträgers auch weitere Administrationsaufgaben für Arbeitgeber übernehmen. Das am Markt häufigste Modell ist das einer Pensionskasse (oder eines Pensionsfonds), die neben der Administration des Versorgungsträgers auch die Verwaltung von Direktzusagen und die Kommunikation mit fremden Versorgungsträgern übernimmt.

3.2.1. Datenquellen des Arbeitgebers

Arbeitgeber verfügen bei aktiv Beschäftigten über Daten zur Versorgung und zum Beschäftigungsverhältnis. In gewissen Umfang verfügt der Arbeitgeber dadurch auch über ein geringes Maß an Daten aus dem privaten Umfeld des Arbeitnehmers. Daten zum Versorgungsbezug fehlen in der Regel, sofern die Versorgung nicht unmittelbar über den Arbeitgeber abgewickelt wird. Insbesondere hinsichtlich des Beschäftigungsverhältnisses verfügt der Arbeitgeber über weitaus mehr Daten als ein Versorgungsträger. Im weitesten Sinne handelt es sich bei den zusätzlichen Daten um Informationen aus dem Beschäftigungsverhältnis, die einem Versorgungsträger üblicherweise nicht vorliegen.

Bei Ausgeschiedenen sind weniger bzw. andere Informationen vorhanden, denn die Informationen, die im Rahmen eines bestehenden Arbeitsverhältnisses gespeichert werden, werden nach Ablauf der Aufbewahrungsfristen zur Einhaltung der Datenschutzanforderungen gelöscht. Entsprechend weniger Informationen stehen nach dem Ausscheiden noch zur Verfügung.

Verfügt ein Arbeitgeber über eine Versorgungsordnung, die weite Teile der Belegschaft auch mit einer Direktzusage versorgt, so verfügt der Arbeitgeber über umfangreiche Informationen zur bAV. Dazu zählen unter anderem auch Informationen

- zum Sterbedatum bei laufenden Leistungen, sofern der Rentner in der Rentenzahlungsphase stirbt,
- zum Sterbedatum von Anwärtern in der aktiven Phase, ggf. auch von ausgeschiedenen Anwärtern vor Rentenbezug,
- über die Hinterbliebenen, wie deren Alter, Adresse, Steuerklasse, Steuer-ID, Sozialversicherungsnummer, Freibeträge etc.

Da also Arbeitgeber mit Direktzusagen über deutlich mehr Daten zur bAV verfügen als Arbeitgeber, die weite Teile der Belegschaft nur über mittelbare Versorgungszusagen versorgen, ist es sinnvoll, in der Folge bei der Taxonomie zwischen

1. Arbeitgebern mit Direktzusagen und
2. Arbeitgebern ohne Direktzusagen

zu unterscheiden.

3.2.2. Datenquellen von externen Versorgungsträgern

Pensionskassen inkl. Zusatzversorgungskassen, Pensionsfonds, Unterstützungskassen und Lebensversicherer als Versorgungsträger verfügen über Informationen zur Altersversorgung von Anwärtern und Rentnern, insbesondere die Höhe der Anwartschaft, Leistungshöhe etc. In geringem Umfang verfügt ein Versorgungsträger auch über Informationen zum Arbeitsverhältnis und in noch geringerem Umfang über private Informationen wie beispielsweise zum Familienstand und zur Gesundheitsdaten, die im Rahmen einer Invaliditätsprüfung abgefragt wurden.

Bei Branchenlösungen und Zusatzversorgungskassen verfügen zudem die durchführenden Versorgungsträger über zusätzliche Daten von Beschäftigungsverläufen der Anwärter innerhalb der Branche. Da dies nur einen Teil der Versorgungsträger betrifft, haben wir dies in der Taxonomie nicht weiter berücksichtigt.

3.2.3. Datenquellen der bAV-Administratoren

bAV-Administratoren sind Dienstleister der Arbeitgeber und verfügen über ähnliche Daten zum Arbeitsverhältnis wie ein Versorgungsträger, da sie nur über die Informationen zum Beschäftigungsverhältnis verfügen, die zur Verwaltung oder Gutachtenerstellung der Zusagen notwendig sind.

Allerdings ist häufig die Situation gegeben, dass ein bAV-Administrator für Arbeitgeber alle (oder zumindest mehrere) Versorgungszusagen verwaltet, d.h.

1. alle (oder mehrere) Direktzusagen sowie
2. alle (oder mehrere) mittelbaren Zusagen, die über vom Arbeitgeber beauftragte Versorgungsträger durchgeführt werden.

Damit sind die Daten, über die ein Administrator verfügt, oftmals umfangreicher als die Daten der Versorgungsträger, da sie alle Zusagen der Beschäftigten umfassen. Da ein bAV-Administrator außerdem die Verwaltung oder Gutachtenerstellung in der Regel für viele Arbeitgeber übernimmt, verfügen große Administrations-Anbieter über die umfangreichsten, durchführungswege- und arbeitgeberübergreifenden Daten.

3.2.4. Datenquellen von hybriden Einrichtungen der bAV

Hybride Einrichtungen der bAV verfügen über Daten wie ein Versorgungsträger und zusätzlich über Daten zur Verwaltung der Zusagen wie bei einem bAV-Administrator. Im Vergleich zu Administratoren verfügen hybride Einrichtungen damit über weitaus mehr Informationen. Dies wird allerdings wiederum dadurch eingeschränkt, dass hybride Einrichtungen ihre Leistung meist nur den Unternehmen einer Branche oder eines Konzerns anbieten.

3.3. Sekundäre Datenquellen

Die Arbeitsgruppe sieht vorrangig bei den folgenden Einrichtungen sekundäre Datenquellen zur bAV:

1. Bundesagentur für Arbeit
2. Deutsche Rentenversicherung
3. Krankenversicherer
4. Finanzämter

Diese Liste ist nicht abschließend.

3.3.1. Datenquellen der Bundesagentur für Arbeit

Die Bundesagentur für Arbeit verarbeitet Daten im Rahmen ihrer gesetzlichen Aufgabenerledigung[40]. Zu den Daten gehören neben Stammdaten unter anderem Daten zur Leistungsgewährung und Gesundheitsdaten[41]. Erfasst werden Daten von Personen und ihren Beschäftigungsverhältnissen aus den letzten 10 Jahren im Rahmen von Arbeitslos- und Arbeitssuchendmeldungen. Löschfristen sowie der Fokus auf nur einen Teil der Erwerbstätigen reduzieren die Möglichkeiten von Analysen im Rahmen des maschinellen Lernens.

3.3.2. Datenquellen der Träger der Deutschen Rentenversicherung

Die Deutsche Rentenversicherung führt für alle durch sie erfassten Erwerbstätigen und Rentner umfangreiche Daten in denen u.a. alle Beschäftigungszeiten und erworbenen Rentenansprüche in Versicherungskonten erfasst werden. Neben den beruflichen Informationen werden eine Vielzahl weiterer Informationen erfasst, so z.B. familiäre Informationen wie die Zahl der Kinder, Finanz- und Zahlungsdaten und Gesundheitsdaten für Rehabilitationsleistungen[42]. Die Erfassung der Gesundheitsdaten ist allerdings beschränkt auf die vorgenannten Rehabilitationsleistungen sowie der Gewährung von Renten wegen (teilweiser) Erwerbsminderung.[43]

3.3.3. Datenquellen von Krankenversicherern

Krankenversicherer verfügen über diverse Informationen, die für Datenanalysen in der bAV relevant sind. So haben Krankenversicherer Kenntnis über das Arbeitsverhältnis und bis zur Beitragsbemessungsgrenze auch Kenntnis über die Höhe des Arbeitseinkommens. Krankenkassen verfügen auch über Informationen darüber, wenn Versicherte mehreren Beschäftigungen nachgehen. Als Einzugsstelle für die Sozialversicherung verfügen Krankenkassen über Informationen zum Beschäftigungsverlauf des Versicherten inkl. der Höhe des Einkommens bis zur Beitragsbemessungsgrenze.

Damit verfügen Krankenkassen über umfangreiche Informationen, die sich zudem von denen der Arbeitgeber und der Versorgungsträger unterscheiden.

Die sicherlich wesentlichste Datenquelle der Krankenversicherer sind die Gesundheitsdaten der Versicherten. In Kombination zu den Beschäftigungsverläufen und der von den Krankenversicherern vorgenommenen Verbeitragung von Leistungen der bAV ließen sich sicherlich aufschlussreiche Analysen durchführen.

Krankenversicherer sind grundsätzlich keine primäre Datenquelle, verfügen aber über umfangreiche Daten in Kombination mit Arbeitsverhältnissen und im Rahmen der Verbeitragung von betrieblicher Altersversorgung in der Leistungsphase. Daher haben wir sie im Rahmen der Taxonomie ebenfalls berücksichtigt und hier beschrieben.

[40] Siehe: https://www.arbeitsagentur.de/datenschutz/datenerhebung abgerufen am 15.09.2023

[41] U.a. für die Betreuung im Reha-Bereich, Begutachtungen oder Stellungnahmen durch den Ärztlichen Dienst der BA, den Medizinischen Dienst der Krankenkassen oder den Berufspsychologischen Service der BA.

[42] Siehe: https://www.deutsche-rentenversicherung.de/SharedDocs/Downloads/DE/Broschueren/national/datenschutz_ihre_daten_und_ihre_rechte.html abgerufen am 20.09.2023

[43] Weitere Informationen zu Daten der Deutschen Rentenversicherung in Abschnitt 3.5.1

3.4. Taxonomie primärer Datenquellen inkl. Krankenversicherer

3.4.1. Motivation und Vorgehen

Dem Pool stehen keine Daten aus primären Datenquellen zur Verfügung. Daher wurde der Ansatz einer theoretischen Taxonomie der Daten verfolgt und auf der Basis des vorhandenen Domänenwissens ausgearbeitet.

Mögliche Attribute aus den Datenquellen wurden beschrieben und Kategorien zugeordnet. Die Kategorien betreffen Daten des Arbeitgebers, Daten zur Person, Daten zum Beschäftigungsverhältnis, Daten zur bAV und Daten zu Hinterbliebenen. Aus Vereinfachungsgründen erfolgt keine Differenzierung, wie die Daten konkret in den Datenbanken abgelegt sind. Stattdessen werden Daten sinnvoll und logisch zusammengefasst. Eine *Adresse* wird in diesem Kontext beispielsweise als ein Attribut erfasst. Es erfolgt keine Differenzierung nach Straße, Hausnummer, Postleitzahl usw. Ebenso werden beispielsweise *Beitragsfreie Zeiten* als ein Attribut erfasst, unabhängig davon, wie die konkrete Ablage der Daten bei der Datenquelle erfolgt.

Insgesamt wurden 121 Attribute über alle Kategorien erfasst. In der folgenden Tabelle findet sich eine Übersicht über die Kategorien inkl. Beispielen. Ebenfalls in der Tabelle enthalten ist die Anzahl der Attribute, die der Pool den Kategorien zugeordnet hat.

Kategorie	Erläuterung	Beispiel(e)	Anz.
Arbeitgeber	Daten des Arbeitgebers	• Name des Arbeitgebers	14
Arbeitgeber bAV	Daten zur bAV des Arbeitgebers	• Anzahl Versorgungswerke	4
Person	Daten der aktiven oder ehemaligen Beschäftigten des Arbeitgebers inklusive der Leistungsempfänger einer beim Arbeitgeber durchgeführten bAV	• Name • Geburtsdatum • SteuerID • Bankverbindung • Behinderung	22
Aktives Arbeitsverhältnis	HR-Daten des Arbeitsverhältnisses eines aktiven Anwärters	• Aktuelles Gehalt • Eintritt in das Unternehmen • Zahl der Krankheitstage	19
Anwärter bAV	Daten zur bAV des Anwärters	• Anzahl Zusagen	4
Zusage	Daten einer Zusage	• Höhe der Entgeltumwandlung	29
Unverfallbar ausgeschiedene Anwärter	Daten zu ausgeschiedenen mit unverfallbarer Anwartschaft	• Austrittsdatum	4
Leistungsempfänger	Daten zu Leistungsempfängern	• Höhe der Leistung	13
Hinterbliebene	Daten zu Hinterbliebenen	• Partner vorhanden • Anzahl Kinder	4
Person Hinterbliebene	Für jeden Hinterbliebenen, Daten zur Person des Hinterbliebenen	• Name • SteuerID	8

In der Taxonomie wurden auch Daten aufgenommen, die nur mittelbar mit einer bAV in Zusammenhang stehen. So hat der Pool beispielsweise Attribute in der Kategorie Arbeitgeber berücksichtigt, die Daten über Krankenquoten beim Arbeitgeber und in der Branche des Arbeitgebers abbilden.

Zu berücksichtigen ist, dass Daten von Beschäftigten nach Ausscheiden in der Regel nach den üblichen Aufbewahrungsfristen im Sinne der DSGVO gelöscht werden. Davon wird nur dann abgewichen, falls die Daten für die bAV benötigt werden. Bei gehaltsabhängigen Bausteinzusagen können z.B. die Gehaltsdaten aufbewahrt werden.

Die in den Kategorien enthaltenen Attribute wurden für die in Abschnitt 3.2 genannten Datenquellen daraufhin analysiert, ob die Datenquelle über Daten

a. sicher verfügt,
b. unter bestimmten Voraussetzungen verfügt oder
c. nicht verfügt.

Sichere Verfügbarkeit bedeutet hier, dass die Datenquelle die Daten verarbeitet hat und daher grundsätzlich über diese Daten verfügt. Falls z.B. eine Zusage eine *flexible Ablaufphase* vorsieht, liegen Daten dazu einem Arbeitgeber grundsätzlich vor. Ob er auf die Daten zugreifen kann, wird durch Taxonomie nicht beantwortet.

Zu den Daten, über die alle Datenquellen verfügen, zählt beispielsweise der Name des Arbeitgebers. Andere Daten, wie z.B. das aktuelle Gehalt eines Beschäftigten sind grundsätzlich nur Arbeitgebern als Datenquelle bekannt. Dagegen verfügen Versorgungsträger nur dann über das Gehalt, falls sie auf der Basis des Gehalts die Höhe eines Beitrags für die bAV des Beschäftigten ermitteln.

Krankenversicherer sind wie oben erläutert, keine primäre Datenquelle. Allerdings verfügen sie aus anderen Gründen über Informationen: So liegen Krankenversicherern Gehaltsinformationen bis zur Höhe der Bemessungsgrenze vor. Zwar ist das konkrete Bruttogehalt einem Krankenversicherer nicht bekannt, dennoch lassen sich daraus Informationen zum finanziellen Hintergrund von Personen ableiten. Da darüber hinaus in der Leistungsphase die Verbeitragung der Leistung erfolgt, verfügen Krankenversicherer ab dem Zeitpunkt der Leistungsphase über Informationen

a. zur Höhe der bAV und
b. zu den verschiedenen Anwartschaften, die ein Versorgungsberechtigter im Laufe des Erwerbslebens aufgebaut hat.

Des Weiteren wurden die Attribute querschnittlich zugeordnet, d.h.

a. zur Person,
b. ihrer Bildung,
c. ihrer Familie,
d. zum finanziellen Hintergrund,
e. der Gesundheit,
f. dem Arbeitsverhältnis,
g. und der bAV.

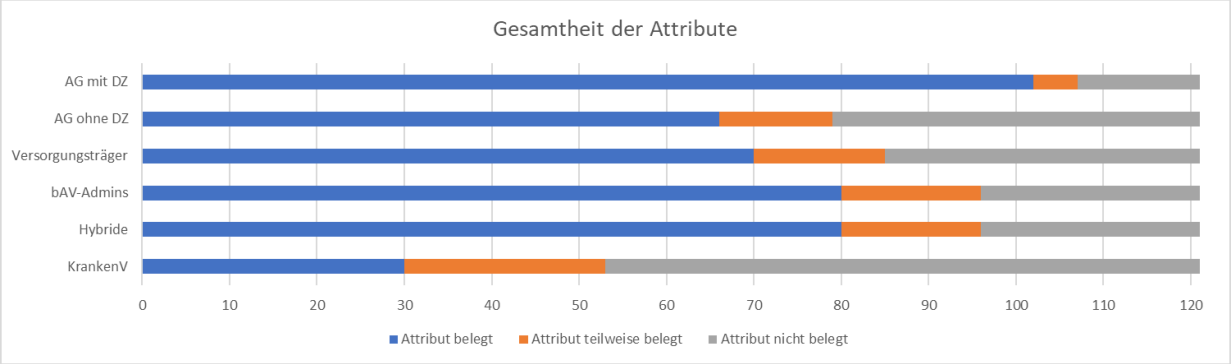
3.4.2. Ergebnis

Es liegen zu Person, Arbeitsverhältnis und Zusage die meisten Attribute vor, gefolgt von Informationen zum Arbeitgeber, Leistungsempfänger und zu Hinterbliebenen.

In der Kombination von Arbeitgeber mit Direktzusage und Krankenversicherer ergibt sich eine nahezu vollständige Abdeckung aller geprüften Attribute. Vergleichbares gilt in der Kombination von

Arbeitgebern ohne Direktzusage, dem durchführenden Versorgungsträger und den Krankenversicherern.

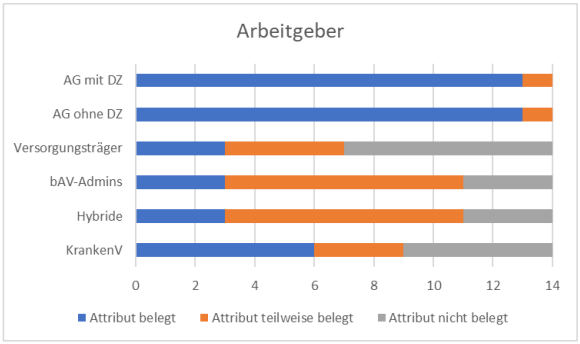
Die folgenden Darstellungen zeigen die Verfügbarkeit von Daten der Datenquellen:

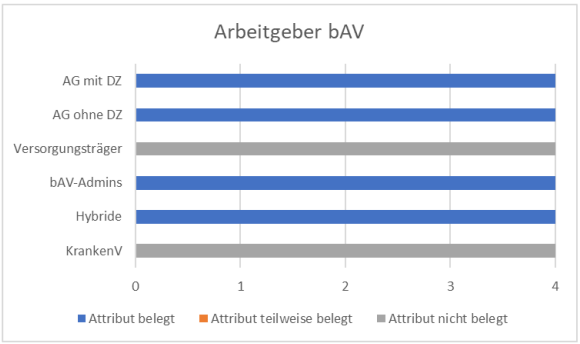


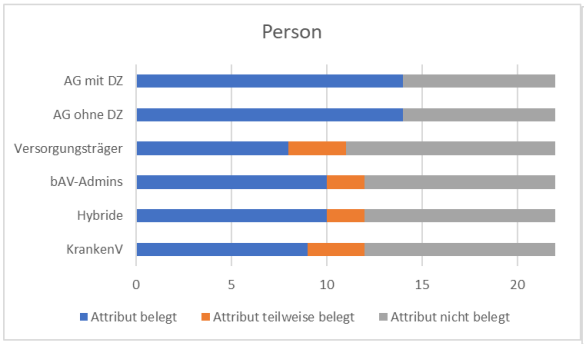
Arbeitgeber mit Direktzusagen verfügen über die größte Abdeckung von Daten. Sie verfügen sowohl über Daten des Arbeitgebers, über umfangreiche Informationen zu den Beschäftigungsverhältnissen und die zugesagte bAV. Vergleichsweise umfangreiche Daten haben auch hybride Einrichtungen der bAV und bAV-Administratoren. Bei diesen primären Datenquellen liegen in der Regel weniger Informationen zum Arbeitsverhältnis vor. Der geringste Umfang an Daten findet sich bei den externen Versorgungsträgern und bei Krankenversicherern. Dafür sind dort umfangreiche Informationen über die Versorgung vorhanden bzw. über die Gesundheitsdaten der Anwärter.

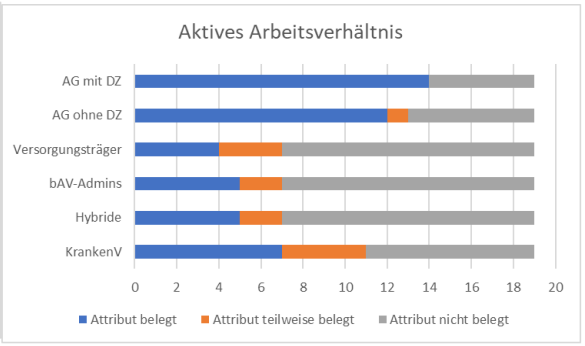
Im folgenden Diagramm werden die Ergebnisse der Taxonomie zur Datenverfügbarkeit in den einzelnen Kategorien dargestellt.

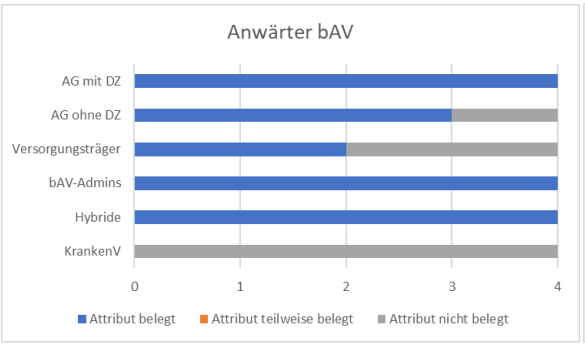
Zur Datenverfügbarkeit bei Hinterbliebenen erklärt sich der Unterschied der Verfügbarkeit zwischen Hinterbliebene und Personen dadurch, dass Hinterbliebene oft erst im Leistungsfall erfasst werden.

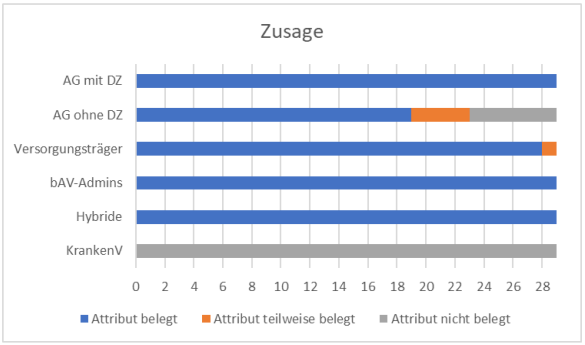

Arbeitgeber
AG mit DZ
AG ohne DZ
Versorgungsträger
bAV-Admins
Hybride
KrankenV
0 2 4 6 8 10 12 14
Attribut belegt
Attribut teilweise belegt
Attribut nicht belegt

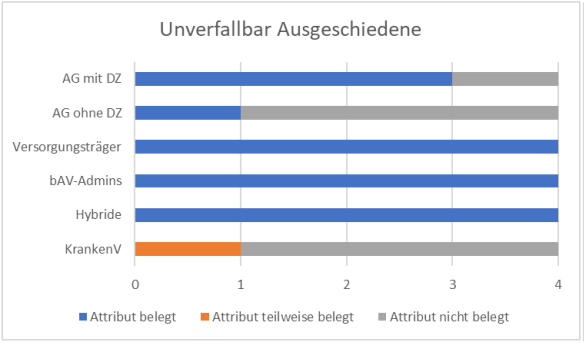

Arbeitgeber bAV
AG mit DZ
AG ohne DZ
Versorgungsträger
bAV-Admins
Hybride
KrankenV
0 1 2 3 4
Attribut belegt
Attribut teilweise belegt
Attribut nicht belegt

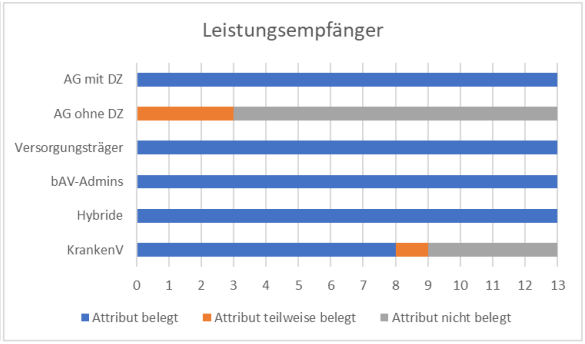

Person
AG mit DZ
AG ohne DZ
Versorgungsträger
bAV-Admins
Hybride
KrankenV
0 5 10 15 20
Attribut belegt
Attribut teilweise belegt
Attribut nicht belegt

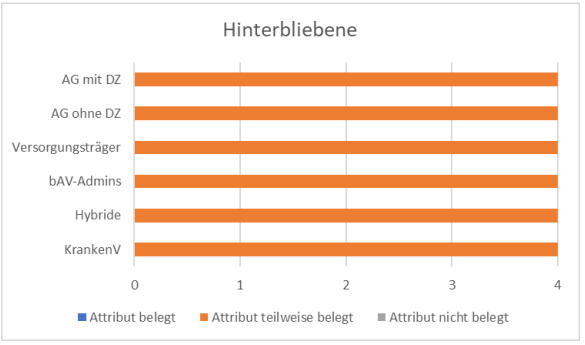

Aktives Arbeitsverhältnis
AG mit DZ
AG ohne DZ
Versorgungsträger
bAV-Admins
Hybride
KrankenV
0 2 4 6 8 10 12 14 16 18 20
Attribut belegt
Attribut teilweise belegt
Attribut nicht belegt

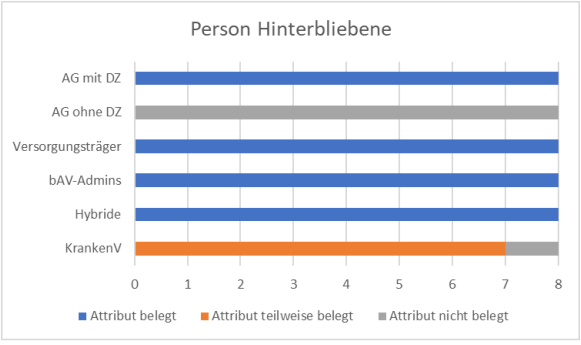

Anwärter bAV
AG mit DZ
AG ohne DZ
Versorgungsträger
bAV-Admins
Hybride
KrankenV
0 1 2 3 4
Attribut belegt
Attribut teilweise belegt
Attribut nicht belegt


Zusage
AG mit DZ
AG ohne DZ
Versorgungsträger
bAV-Admins
Hybride
KrankenV
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28
Attribut belegt
Attribut teilweise belegt
Attribut nicht belegt


Unverfallbar Ausgeschiedene
AG mit DZ
AG ohne DZ
Versorgungsträger
bAV-Admins
Hybride
KrankenV
0 1 2 3 4
Attribut belegt
Attribut teilweise belegt
Attribut nicht belegt


Leistungsempfänger
AG mit DZ
AG ohne DZ
Versorgungsträger
bAV-Admins
Hybride
KrankenV
0 1 2 3 4 5 6 7 8 9 10 11 12 13
Attribut belegt
Attribut teilweise belegt
Attribut nicht belegt


Hinterbliebene
AG mit DZ
AG ohne DZ
Versorgungsträger
bAV-Admins
Hybride
KrankenV
0 1 2 3 4
Attribut belegt
Attribut teilweise belegt
Attribut nicht belegt


Person Hinterbliebene
AG mit DZ
AG ohne DZ
Versorgungsträger
bAV-Admins
Hybride
KrankenV
0 1 2 3 4 5 6 7 8
Attribut belegt
Attribut teilweise belegt
Attribut nicht belegt

3.5. Anreichernde Datenquellen

3.5.1. Forschungsdatenzentrum der Rentenversicherung

Das Forschungsdatenzentrum der Rentenversicherung (FDZ-RV) stellt Daten sowohl für die interessierte Öffentlichkeit[44] als auch für die wissenschaftliche Forschung[45] bereit. Bei den Daten für die interessierte Öffentlichkeit handelt es sich weitestgehend nicht um Daten im eigentlichen Sinne, sondern zumeist nur um die Beschreibung der Daten sowie Statistiken, die auf der Basis dieser Daten erstellt werden. Exemplarisch sei hier der Versichertenrentenzugang 2018 erwähnt, der aktuell auf der Homepage der FDZ-RV bereitgestellt wird. Die Daten enthalten umfangreiche und detaillierte soziodemografische Merkmale wie Geschlecht, Geburtsjahr, Familienstand, Nationalität, Wohnort (Bundesland, Arbeitsmarktregion, Kreis- und Regionstyp) und rentenspezifische Merkmale wie aktueller/erstmaliger Rentenbeginn, Höhe und Zusammensetzung der Anwartschaften, rentenrechtliche Zeiten, Fremdrenten, Entgelte, Leistungsart, Rentenart, Rentenzahlbetrag. Bereitgestellt werden unter anderem eine ausführliche Datensatzbeschreibung sowie der Statistikband „Rente 2018“, der auf Basis einer 1% Zufallsstichprobe mit 203.357 anonymisierten Datensätzen erstellt wurde. Die Daten selbst werden nicht, auch nicht anonymisiert, zur Verfügung gestellt.

Für die interessierte Öffentlichkeit gibt es des Weiteren sogenannte Regionalfiles, die als Open Data bereitgestellt werden. Dabei handelt es sich um aggregierte Daten für Landkreise und Städte, die unter anderem die Anzahl der Versicherten, Rentner und Rentenzugänge oder Abgänge je Kreis oder Stadt enthalten. Die Regionalfiles enthalten keine Daten, die im Rahmen aktuarieller Analysen genutzt werden können.

Etwas anderes gilt für die sogenannten Scientific Use Files, die interessierten Wissenschaftlern zu Forschungszwecken bereitgestellt werden. Dabei handelt es sich sowohl um Längsschnittdaten (also ein Berichtzeitraum über mehrere Jahre) als auch um Querschnittdaten (Stichtage oder Berichtsjahre). Im Bereich der Querschnittdaten können Forscher derzeit auf

- den Versichertenrentenbestand 2021,
- den Versichertenrentenzugang 2021
- und die aktiv Versicherten 2020

zugreifen. Im Bereich der Längsschnittdaten stehen Daten zu

- abgeschlossene Rehabilitation im Versicherungsverlauf 2013 bis 2020,
- Versicherungskontenstichprobe 2019 (Biografiedaten zu Versicherten)
- und vollendete Versichertenleben 2020 (Biografiedaten zu Vollendeten Versichertenleben)

zur Verfügung.

Der Öffentlichkeit stehen zu diesen Daten nur die Datensatzbeschreibungen und die Statistikbände zur Verfügung. Bei der Durchsicht der Datensatzbeschreibungen wird deutlich, dass die Daten äußerst relevante Informationen für Data-Science-Analysen der bAV enthalten. Beispielhaft sei hier auf die Datensatzbeschreibung der abgeschlossenen Rehabilitationen verwiesen. Dort finden sich neben den Daten zur Diagnose und umfangreichen Informationen zur Rehabilitation auch Informationen zu Geburtsjahr und -monat, Schulbildung, beruflichen Ausbildung, zum Beruf, der Stellung im Beruf und zum Wohnsitz. Allerdings stehen diese Daten Aktuaren außerhalb von Forschungseinrichtungen nicht zur Verfügung.

[44] Siehe: https://www.eservice-drv.de/FdzPortalWeb/dispcontent.do?id=main_fdz_lehre&chmenu=ispvwNavEntriesByHierarchy62 abgerufen am 01.01.2024

[45] Siehe: https://www.eservice-drv.de/FdzPortalWeb/dispcontent.do?id=main_fdz_forschung&chmenu=ispvwNavEntriesByHierarchy27 abgerufen am 01.01.2024

Dass umfangreiche Datenanalysen möglich sind, wird beispielsweise an den Auswertungen von Haan et al.[46] deutlich, die Lebenserwartungen westdeutscher Arbeitnehmer in Abhängigkeit von Geburtsjahrgang und Lebenslohneinkommen darstellen.

3.5.2. SOEP-RV

SOEP-RV ist ein gemeinsames Projekt des Forschungsdatenzentrums der Rentenversicherung mit dem Deutschen Institut für Wirtschaftsforschung (DIW) und wird unter anderem über das Forschungsdatenzentrum der Rentenversicherung bereitgestellt[47]. Im Projekt werden Daten des **S**ozio-**oe**konomischen **P**anels (SOEP)[48] mit anonymisierten FDZ-RV-Daten verknüpft.

Bei SOEP handelt es sich um eine repräsentative Wiederholungsbefragung von Privathaushalten in Deutschland, die Daten aus jährlichen Befragungen von 30.000 Menschen in 15.000 Haushalten enthält. Im Rahmen des Projektes SOEP-RV werden diese Daten mit ihren Versicherungskonten oder Versicherungsrenten verknüpft.

Die Daten des SOEP enthalten Informationen zu Bildung, dem Arbeitgeber, der Art, Zeiten und Umfang der Beschäftigung, Weg zur Arbeit, Einkommen, Krankheiten und Wohnsituation sowie subjektive Merkmale über die Lebenszufriedenheit. Daten zu Einkommen werden detailliert aufgeschlüsselt und enthalten z.B. Informationen darüber, ob ein 13. oder 14. Monatsgehalt gezahlt wird oder ob Fahrtkosten erstattet werden. Ebenso werden detailliert die Form und der Umfang der privaten, gesetzlichen oder betrieblichen Altersversorgung abgefragt. Im Rahmen dieser Langzeitbefragungen werden Personen, die in einen befragten Haushalt hinzukommen (Kinder, Partner) in die Befragung aufgenommen. Personen, die den Haushalt verlassen (Scheidung) werden weiterhin befragt.

Die Daten enthalten äußerst relevante Informationen für Data-Science-Analysen der bAV, stehen aber Aktuaren außerhalb von Forschungseinrichtungen nicht zur Verfügung.

3.5.3. SHARE-RV

SHARE-RV ist ein Kooperationsprojekt zwischen dem Munich Center for the Economics of Aging (MEA) des Max-Planck-Instituts für Sozialrecht und Sozialpolitik und dem Forschungsdatenzentrum der Deutschen Rentenversicherung und wird ähnlich wie SOEP-RV unter anderem über das Forschungsdatenzentrum der Rentenversicherung bereitgestellt[49]. Das Projekt befragt regelmäßig Menschen im Alter über 50 länderübergreifend in Europa und in Israel.[50]

Vergleichbar mit SOEP werden Menschen wiederholt befragt und Daten zu Gesundheit, Einkommen, Vermögen, Konsum, Bildung, Wohnsituation, Arbeit und ehrenamtliche Tätigkeiten erfasst. Bei der Befragung liegt unter anderem ein Fokus auf Gesundheitszustand, Gesundheitsverhalten und körperlichen Indikatoren wie der Greifkraft, sowie auf die sozialen Netze wie der Familie und der psychischen Gesundheit und die Lebenszufriedenheit. Ebenfalls werden detailliert gewöhnliche (tägliche) Tätigkeiten wie außergewöhnliche Tätigkeiten abgefragt, um ein möglichst genaues Bild der befragten Person zu erhalten.

---

[46] Haan, P, Kemptner, D. & Lüthen, H. (2019). Besserverdienende profitieren in der Rentenversicherung zunehmend von höherer Lebenserwartung, DIW Wochenbericht 23

[47] Siehe https://www.eservice-drv.de/FdzPortalWeb/dispcontent.do?id=main_fdz_soep-rv&chmenu=ispvwNavEntriesByHierarchy104 abgerufen am 01.01.2024

[48] Siehe https://www.diw.de/de/diw_01.c.678568.de/forschungsdatenzentrum_soep.html abgerufen am 02.01.2024

[49] Siehe https://www.eservice-drv.de/FdzPortalWeb/dispcontent.do?id=main_fdz_share-rv&chmenu=ispvwNavEntriesByHierarchy97 abgerufen am 03.01.2024

[50] Siehe https://www.share-eric.eu/ abgerufen am 03.01.2024

Ziel der Studie ist es, Daten zur Erforschung von Alterungsprozessen in den untersuchten Ländern bereitzustellen. Im Rahmen des Projektes SHARE-RV werden Daten von Personen mit Anwartschaften oder Renten der Deutschen Rentenversicherung mit ihren Versicherungskonten oder Versicherungsrenten verknüpft.

Die Daten enthalten sehr relevante Informationen für Data-Science-Analysen der bAV, stehen aber Aktuaren außerhalb von Forschungseinrichtungen nicht zur Verfügung.

3.5.4. NHANES Datensatz

"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.[51]

Die ersten Datenerhebungen des Programms erfolgten bereits im Jahr 1959 (damals noch unter dem Namen „NHES“) und werden seitdem bis heute regelmäßig durchgeführt. Die Datenbasis wird also laufend erweitert. Veröffentlicht werden dabei nur die Daten von Erwachsenen ab dem 18. Lebensjahr.

Das Programm lässt sich in mehrere Phasen einteilen, wobei die genauen Inhalte der Befragungen und Untersuchungen über die Jahre mehrmals angepasst wurden. Grundsätzlich lassen sich die erhobenen Informationen einteilen in fünf Klassen: Demographie, Ernährung, Labor, Untersuchung und Fragebogen. Insgesamt ergibt sich daraus eine sehr reichhaltige Datenquelle mit mehreren Tausend gemessenen Variablen von Zehntausenden US-Bürgern.

Die DAV hat die Aufbereitung eines Teils dieser Daten im Detail übernommen und veröffentlicht (siehe Ergebnisbericht Sterblichkeitsanalyse des NHANES-Datensatz mit Data Science-Methoden“ vom 21.9.2022 und zugehörigen R-Code, für den dort der Zugriff beschrieben ist). Es handelt sich dabei um die Daten aus den Programm-Phasen „NHANES III“ und „NHANES Continuous“, wobei das erste Programm von 1988 bis 1994 und das zweite von 1999 bis heute durchgeführt wurde. Die Aufbereitung ergibt einen Datensatz mit etwa 65.000 Personen und 70 Variablen. Dieser Datensatz lässt sich auch noch um eine Vielzahl an weiteren Variablen erweitern.

In der aktuellen Programm-Phase („NHANES Continuous“) werden in zweijährigen Zyklen ca. 7.000 Personen aller Altersgruppen zu Hause befragt, wovon ca. 5.000 die Gesundheitsuntersuchungskomponente der Umfrage absolvieren. Jede Person wird dabei allerdings nur einmal befragt und nicht über die Jahre begleitet. Lediglich Todeszeitpunkt und Todesursache lassen sich nachträglich über sogenannte Linked Mortality Files anfügen. Diese Dateien beruhen auf Informationen des National Death Indexes und werden vom CDC veröffentlicht.

Zusätzlich wurde eine Langzeitstudie in der Programm-Phase „NHEFS“ (NHANES I Epidemiologic Follow-up Study) durchgeführt. Dabei wurden 14.407 Studienteilnehmer der NHANES I Kohorte (1971-1975) über mehrere Jahre begleitet. Diese Personen wurden viermal befragt in den Jahren 1982/84, 1986, 1987 und 1992. Es lassen sich also insbesondere veränderte Lebensumstände und Krankheitsentwicklungen analysieren, da beispielsweise Merkmale wie Raucher, Todesursache, Alkoholkonsum, Einschränkungen im Alltag, Krankenhausbesuche, Bildungsgrad, Einkommen und Familienstand vorhanden sind. Der Zugang zu diesen Daten ist ähnlich zu den vom DAV-Ergebnisbericht betrachteten Daten. Allerdings wurden die Befragungen nicht immer standardisiert durchgeführt, was einen nicht zu unterschätzenden manuellen Aufwand beim Erstellen von Datensätzen verlangt.

Insgesamt ist festzuhalten, dass „NHANES Continuous“, auch im Vergleich zur Langzeitstudie „NHEFS“, als besonders wertvolle Datenquelle einzuschätzen ist, da diese Daten laufend erweitert und standardisiert erhoben werden.

Bei der Durchführung von Analysen auf den NHANES-Daten ist natürlich darauf zu achten, inwieweit die Ergebnisse auf Deutschland übertragen werden können.

[51] Umfangreiche Informationen unter www.cdc.gov/nchs/nhanes

4. Mögliche Anwendungsbeispiele

4.1. Einführung

In diesem Abschnitt werden mögliche Anwendungsbeispiele für die bAV ohne Anspruch auf Vollständigkeit aufgezeigt. Der Pool verzichtet darauf, Anwendungen zu beschreiben, die im Rahmen eines normalen Geschäftsbetriebs außerhalb der bAV eingesetzt werden können und die sich nicht wesentlich von anderen Sparten unterscheiden. Dazu zählen beispielsweise Verfahren zur Text- und Spracherkennung, die beispielsweise zur Klassifizierung und Auswertung von Dokumenten eingesetzt werden und sich zwischen den Sparten nur durch das konkrete Training der Verfahren unterscheiden.

Der Pool verzichtet ebenfalls darauf, auf Anwendungen einzugehen, bei denen Verfahren zur Vorhersage von Verhalten eingesetzt werden. Diese können beispielsweise eingesetzt werden zur Vorhersage von Emotionen und Stimmungen in der Kommunikation mit Versorgungsberechtigten oder Arbeitgebern oder zur Ansprache von potenziellen Anwärtern im Rahmen von vertrieblichen Marketingmaßnahmen.

4.2. Biometrische Rechnungsgrundlagen

Für versicherungsmathematische Bewertungen in der bAV kommen neben dem Rechnungszinssatz den biometrischen Rechnungsgrundlagen eine maßgebliche Bedeutung zu. Die Herleitung, Überprüfung und Verwendung von angemessenen biometrischen Rechnungsgrundlagen gehört daher zu den Kernaufgaben eines Aktuars. Dies gilt gleichermaßen für die Bewertung von Pensionsverpflichtungen bei Direktzusagen sowie für die Bewertung von Deckungsrückstellungen einer Einrichtung der betrieblichen Altersversorgung.

Die biometrischen Rechnungsgrundlagen umfassen z.B. Informationen über

- Sterblichkeit,
- Invalidisierung,
- Verheiratungswahrscheinlichkeit,
- Altersunterschiede bei Ehegatten,
- Inanspruchnahme von Auszahlungsoptionen,
- Renteneintrittsalter,
- und Fluktuation.

Bei der Herleitung und Validierung dieser Rechnungsgrundlagen kommen derzeit üblicherweise mathematisch-statistische Testverfahren zum Einsatz.[52] [53]

Im Rahmen der Herleitung der Heubeck Richttafeln 2018 G wurden bei den Sterbewahrscheinlichkeiten erstmalig auch die sozioökonomischen Lebensumstände über einen einheitlichen Faktor berücksichtigt. Würde bei der Untersuchung der biometrischen Rechnungsgrundlagen neben dem Geburtsjahrgang, Alter und Geschlecht noch weitere Merkmale wie Wohnort, Arbeitsbedingungen,

[52] Fachgrundsatz der DAV „Biometrische Rechnungsgrundlagen bei Pensionskassen und Pensionsfonds“ in der Fassung vom 28.01.2019
https://aktuar.de/unsere-themen/fachgrundsaetze-oeffentlich/2019-01-28_Hinweis-Biometrische-Rechnungsgrundlagen-PKs-PFs.pdf

[53] BMF-Schreiben vom 9.12.2011, Anerkennung unternehmensspezifischer und modifizierter biometrischer Rechnungsgrundlagen
https://www.bundesfinanzministerium.de/Content/DE/Standardartikel/Themen/Steuern/Steuerliche_Themengebiete/Altersvorsorge/2011-12-09-Betriebliche-Altersversorgung-Bewertung-Pensionsverpflichtungen.pdf?__blob=publicationFile&v=3

vertragliche Rahmenbedingungen der Arbeitsverhältnisse und weitere persönliche Daten berücksichtigt, so wäre auf Grund der gestiegenen Komplexität dies auch ein Anwendungsgebiet für Data Science in der bAV.

Sofern rechtliche Vorgaben oder der Schutz der Mitarbeiterdaten die Verwendung von zusätzlichen Merkmalen für versicherungsmathematische Jahresabschlussbewertungen nicht zulassen, so wäre ein Ansatz zusätzlicher Merkmale und der daraus abgeleiteten Erkenntnisse für ausschließlich interne Zwecke z.B. für das Risikomanagement denkbar.

Ein vergleichbares Anwendungsbeispiel in der Lebensversicherung ist die Analyse und Vorhersage von Stornowahrscheinlichkeiten oder die Modellierung des Gesundheitszustandes der Versicherten.[54]

4.3. Bestandsverdichtungen

"Mit wachsenden Anforderungen an stochastische Bewertungen der großen Lebensversicherungsportfolien stoßen Unternehmen auf ein Laufzeitproblem. Es sollen grundsätzlich Millionen Verträge unter Tausenden Zinsszenarien über mehrere Jahrzehnte und unter zahlreichen Annahmen projiziert werden. Ziel einer Bestandsverdichtung ist die Erzeugung eines möglichst kleinen Teilbestandes, der die gleichen Eigenschaften besitzt wie der Originalbestand." (Ergebnisbericht „Big Data in der Lebensversicherung")

In der bAV stehen EbAVs grundsätzlich vor ähnlichen Herausforderungen wie die Lebensversicherungen. Allerdings ist aufgrund des regulatorischen Umfeldes und der Größe der Versicherungsbestände der Bedarf an einer Verdichtung von Beständen weniger praxisrelevant. Nichtsdestotrotz kann im Rahmen von ALM-Studien (z.B. zur Überprüfung der Strategischen Asset Allokation) eine Bestandsverdichtung erforderlich werden, um eine stochastische Bewertung vornehmen zu können.

Im Rahmen von versicherungsmathematischen Bewertungen für den Jahresabschluss sind aufgrund des Prinzips der Einzelbewertung Bestandsverdichtungen nicht zulässig.

4.4. Validierung von Bewertungsergebnissen

Im Rahmen von versicherungsmathematischen Bewertungen werden mit Hilfe von Personenstammdaten, Bilanzinformationen, Informationen zur Zusage oder des Versicherungsvertrages und der gewählten Rechnungsgrundlagen eine Vielzahl an Ergebnisdaten erzeugt. Es liegt in der Verantwortung des jeweiligen Aktuars diese Ergebnisse auf Fehler und unplausible Entwicklungen im Vergleich zu vorherigen Bewertungen zu überprüfen.

Bei der einzelvertraglichen Auswertung von zukünftigen Zahlungsströmen für Beiträge und Leistungen und deren Vergleich mit der korrespondierenden vorherigen Bewertung lassen sich mit Hilfe von Verfahren zur Mustererkennung ggf. Ausreißer identifizieren, die zusätzliche Hinweise auf unplausible Bewertungsergebnisse liefern.

Im Gegensatz zu vielen anderen Anwendungsbeispielen besteht hier kein Mangel an Daten oder zusätzlichen rechtlichen Hürden bei der Datenverarbeitung. Schließlich stehen die Daten für die eigentliche Bewertung zur Verfügung, die daraus abgeleiteten Ergebnisdaten wurden selbst erzeugt und der Zweck der Datenverarbeitung besteht unverändert fort.

[54] Ergebnisbericht des Ausschusses Lebensversicherung „Big Data in der Lebensversicherung" in der Fassung vom 19.09.2019
https://aktuar.de/unsere-themen/fachgrundsaetze-oeffentlich/2019-09-19_Ergebnisbericht_Big-Data-Leben_final.pdf

4.5. Auswertung der Ergebnisse stochastischer Projektionsmodelle

Zunehmend bilden unternehmensspezifische Projektionsmodelle das Fundament für Entscheidungsfindungen und für ein effizientes Risikomanagement. Diese Modelle basieren auf einer Vielzahl von Kapitalmarktszenarien und berücksichtigen die Unternehmensdaten sowie die Strategie(n) des Unternehmens (Managementregeln). Mit Hilfe dieser Informationen werden viele Datensätze erzeugt, um kurz- und langfristige Auswirkungen auf Unternehmenskennzahlen zu analysieren. Dabei ist zu beobachten, dass die Modelle komplexer werden, was die Validierung des Modells und die Analyse der Ergebnisse zeitaufwendiger macht.

Aus Sicht des Pools sind Data-Science-Methoden für die Ergebnisanalyse anwendbar – schließlich lassen sich anhand der Anzahl verwendeter Kapitalmarktszenarien genügend Daten durch das Projektionsmodell erzeugen. Im Folgenden werden drei Fragestellungen aus der Praxis vorgestellt und skizziert, wie diese mit Hilfe von Data-Science-Methoden beantwortet werden könnten.

Zum leichteren Verständnis der Fragestellungen stellen wir uns beispielhaft als Unternehmen eine Pensionskasse vor, die im Rahmen ihrer eigenen Risikobeurteilung nach § 234d VAG oder für die Festlegung der strategischen Asset Allokation eine ALM-Studie durchgeführt hat. Im Rahmen dieser Studie wurden 10.000 Kapitalmarktszenarien erzeugt und pro Szenario die Auswirkungen auf die Bilanz und die GuV über einen Betrachtungszeitraum von 15 Jahren berechnet.

Fragestellung 1: Welches sind die wesentlichen Risiken für das Unternehmen?

Während einige wesentlichen Risiken offensichtlich sind (z.B. alle Szenarien, in denen sich ein Niedrigzinsumfeld realisiert), treten andere wesentlichen Risiken selten auf oder sind nur als Kombination von mehreren einzelnen Faktoren relevant (z.B. hohe Zinsen bei gleichzeitiger hoher Inanspruchnahme von Kapitalwahlrechten). Hier können Verfahren der Mustererkennung in Cashflows oder Kennzahlen dazu beitragen, den Leser der Studie auf weitere „Risiko-Cluster“ in den untersuchten Kapitalmarktszenarien aufmerksam zu machen.[55] Das Erkennen von Risiken setzt natürlich voraus, dass die jeweiligen Einflussgrößen auch im Projektionsmodell abgebildet werden.

Indem kritische Konstellationen von Risikofaktoren mittels Data-Science-Verfahren identifiziert werden, können Ergebnisse des Modells besser erklärt werden. Dies erhöht das Verständnis für das Modell und für die Risiken des Unternehmens.

Fragestellung 2: Wie verändert sich die Interpretation der Ergebnisse des Projektionsmodells, wenn im Zeitverlauf die gewählten Annahmen durch die tatsächliche Entwicklung überholt werden?

Nehmen wir an, seit der ALM-Studie ist ein Jahr vergangenen, die tatsächliche Entwicklung der Kapitalmärkte in diesem Jahr steht fest und eine weitere Bilanz und GuV der Pensionskasse liegt vor. Von den ursprünglich 10.000 simulierten Kapitalmarktszenarien sind nun z.B. 8.000 von der Realität überholt, da sich die tatsächliche Entwicklung signifikant von der erwarteten Entwicklung dieser Szenarien unterscheidet. Auf Grundlage der Ergebnisse der verbleibenden 2.000 Szenarien lassen sich aber weiterhin belastbare Prognosen der Unternehmensentwicklung vornehmen, ohne eine erneute ALM-Studie durchführen zu müssen. Für die Auswahl der geeigneten Szenarien kann wiederum auf Verfahren der Mustererkennung zurückgegriffen werden. Die Anzahl der durch die Algorithmen gefundenen geeigneten Szenarien, wäre somit gleichzeitig ein Indikator für die im Zeitverlauf (abnehmende) Güte der Studienergebnisse.

Fragestellung 3: Können für ausgewählte Risikofaktoren einfache Prognosen außerhalb der berechneten Szenarien vorgenommen werden?

Das Modell hinter der ALM-Studie ist meist ein komplexes und in der Handhabung (rechen-)zeitaufwendiges Modell, welches nicht kurzfristig für einzelne Fragestellungen eingesetzt werden kann. Erstrebenswert wäre die Nachbildung des vollständigen Modells durch ein Data-Science-

[55] Siehe dazu beispielsweise auch Pierre Joos, Machine Learning in Life Insurance: Searching for Patterns in Cash Flows, https://www.actuview.com/, zuletzt abgerufen: 25.05.2023.

Modell, um ad-hoc und ohne großen Aufwand den Einfluss der Änderung einzelner Merkmale bestimmen zu können.

Hierfür muss das Data-Science-Modell anhand der 10.000 berechneten Simulationen kalibriert werden. Für robuste Prognoseergebnisse empfiehlt es sich, sich auf die ausgewählten Merkmale zu fokussieren, die später untersucht werden sollen (z.B. die Annahme für langfristige Inflation).

Im kalibrierten vereinfachten Modell wäre dann z.B. die Inflationsannahme eine Eingabegröße und als Ergebnis wird die geänderte Eigenmittelquote in fünf Jahren betrachtet.

Neues Anwendungsgebiet: Reine Beitragszusage

Das Gesetz zur Stärkung der betrieblichen Altersversorgung und zur Änderung anderer Gesetze (Betriebsrentenstärkungsgesetz) vom 17. August 2017 führte die reine Beitragszusage in die bAV Deutschlands ein. Bei dieser Form der bAV liegt ein wesentliches Risiko in den Rentenkürzungen. Der Gesetzgeber hat hierfür das Prüfkriterium des Kapitaldeckungsgrades eingeführt. So muss die durchführende Einrichtung gewährleisten, dass der Kapitaldeckungsgrad jederzeit zwischen 100 % und 125 % liegt. Zudem sind die Sozialpartner an der Durchführung und Steuerung zu beteiligen und einzubinden. Hierfür ist ein effektives Risikomanagement erforderlich, welches Verfahren zur Messung, Überwachung und Steuerung bereitstellen sollte. Um den Anforderungen gerecht zu werden, kann es ein Mittel der Wahl sein, Prognosen vor allem über kurze Zeiträume wie z.B. monatliche Entwicklungen zu erstellen. Hierfür bieten sich Data-Science-Methoden an, da klassische Projektionsmodelle wie oben beschrieben zu zeitaufwendig und komplex für diese Zwecke sein könnten.

5. Fazit

Um Data Science effektiv betreiben zu können, ist es unerlässlich, über eine ausreichende Menge an qualitativ hochwertigen Daten zu verfügen. Aktuare im Bereich der betrieblichen Altersversorgung profitieren hier von ihrer Expertise in der bAV und sind häufig auch mit dem unternehmensinternen Management der Daten vertraut. Dies ist notwendig, da es einerseits gilt, einen umfangreichen rechtlichen Rahmen einzuhalten wie der Abschnitt Rechtliche Rahmenbedingungen zeigt. Andererseits sind die zur Verfügung stehenden Daten mitunter auf einige wenige Merkmale beschränkt – möglicherweise zu wenige für umfassende Analysen oder um Data-Science-Modelle anzuwenden.

Es ist wichtig zu erkennen, dass oft mehr Daten vorhanden sind, als zunächst angenommen wird. In diesem Bericht wurde eine Vielzahl von Datenquellen vorgestellt, die der bAV zugeordnet werden können. Nicht alle dieser Datenquellen stehen für Data-Science-Analysen der bAV unmittelbar zur Verfügung. Hinsichtlich der Datenverfügbarkeit wurde jedoch eine Taxonomie primärer Datenquellen inkl. Krankenversicherer entwickelt und veranschaulicht.

Als weitere wertvolle Ressource wurden selbst generierte Daten identifiziert, die in der aktuariellen Arbeit entstehen. Diese Daten können ohne rechtliche Hürden genutzt werden, zum Beispiel zur Validierung von Bewertungsergebnissen oder zur Auswertung der Ergebnisse stochastischer Projektionsmodelle.

Aufbereitete Datensätze für aktuarielle Zwecke zu verwenden, erfordert bisher viel Zeit und Expertenwissen. Data Science kann hier Abhilfe schaffen und der Bericht zeigt beispielhafte Anwendungen auf:

- Biometrische Rechnungsgrundlagen können aus umfangreicheren Merkmalen abgeleitet werden oder
- Bestandsverdichtungen für Projektionsrechnungen durchgeführt werden.
- Kurzfristprognosen könnten mit einfacheren Modellen schneller erstellt werden als mit komplexeren Projektionsmodellen.

Abschließend verbleibt den Akteuren die Aufgabe, darüber nachzudenken, wie sie ihren Datenschatz verbessern können, um ihr Geschäft in Zukunft mit Hilfe von Data Science besser zu verstehen und davon zu profitieren.