DAV
Deutsche
Aktuarvereinigung e.V.

Ergebnisbericht des Ausschusses Krankenversicherung

Künstliche Intelligenz und moderne Methoden in der Krankenversicherung

Köln, 2. Oktober 2026

Präambel

Die Arbeitsgruppe Künstliche Intelligenz und moderne Methoden in der Krankenversicherung des Ausschusses Krankenversicherung der Deutschen Aktuarvereinigung e. V. (DAV) hat den vorliegenden Ergebnisbericht erstellt.[1]

Anwendungsbereich

Der Ergebnisbericht betrifft Aktuarinnen und Aktuare[2] bei der Ausführung aktuarieller Aufgaben, schwerpunktmäßig im Bereich der Krankenversicherung.

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 zur Anwendung von künstlicher Intelligenz und weiteren modernen Methoden in der aktuariellen Arbeit der Krankenversicherung. Es werden sowohl rechtliche Rahmenbedingungen betrachtet und Anforderungen an Aktuarinnen und Aktuare abgeleitet als auch praktische Anwendungsfälle skizziert und Beispiele zur praktischen Nutzung zur Verfügung gestellt.[4]

Schlagworte

Krankenversicherung, Künstliche Intelligenz, Maschinelles Lernen, Large Language Models, EU AI Act, DORA, Governance

Verabschiedung

Dieser Ergebnisbericht ist durch die Ausschüsse Krankenversicherung und Actuarial Data Science/AI am 2. Oktober 2026 verabschiedet worden.

---

[1] Der Ausschuss dankt der Arbeitsgruppe Künstliche Intelligenz und moderne Methoden in der Krankenversicherung ausdrücklich für die geleistete Arbeit, namentlich Grit Läuter-Lüttig (Leitung), Dr. Simon Hatzesberger (Co-Leitung) bzw. Axel Kaiser (Co-Leitung), Kinga Böhm, Dr. Fabian Bohnert, Christian Hartmann, Daniel Hasler, Sebastian Heinrichs, Stephanie Heitmüller, Dr. Marius Hermanns, Shobiga Jeyadevan, Dr. Daniel König, Friedrich Loser, Dr. Christoph Nell, Marco Nickel, Stefan Nohl, Dr. Jens Piontkowski, Stephan Rudolph, Philip Schallert und Andreas Schneider.

[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 bzw. Aktuarin DAV getroffen werden.

[4] Bei der Erstellung dieses Berichts wurden Large Language Models unterstützend zur Recherche und zur Formulierung einzelner Passagen eingesetzt. Alle Ideen, Inhalte, Bewertungen und Schlussfolgerungen stammen von den Mitgliedern der Arbeitsgruppe, die sie in vollem Umfang verantwortet.

This abstract summarises the report on findings „Künstliche Intelligenz und moderne Methoden in der Krankenversicherung“ which was approved by the DAV Health Insurance committee on 02.10.2026.

Artificial Intelligence and Modern Methods in an Actuarial Health Insurance Context

This report addresses the use of artificial intelligence and other modern methods in actuarial work in health insurance. Chapter 2 defines the underlying terms: AI systems, modern methods, automation tools and AI agents. Chapter 3 sets out the regulatory framework, covering the classification of AI systems under the EU AI Act, the roles of provider and deployer, the requirements for high-risk systems, the ICT governance obligations arising from DORA, and the resulting requirements for data, models and processes. Chapter 4 maps concrete opportunities along the premium adjustment process of German private health insurance and in the analysis of in-force portfolio data. Chapter 5 presents practical examples developed by the working group, among them a tool for capping premium increases, an LLM-supported validation of technical calculation bases, an approach to using diagnosis data in claims forecasting, and two publicly available analysis notebooks. Chapter 6 describes the responsibilities of the actuarial function and the particular role of the Appointed Actuary. Two annexes introduce the underlying methods of machine learning and of generative artificial intelligence. The report concludes that the risks associated with these methods can be controlled through clearly defined processes, consistent documentation, sound governance and targeted human oversight, and that they can therefore be applied responsibly across a broad range of actuarial use cases.

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 7
1.1. Relevanz der Fragestellung 7
1.2. Zielsetzung 7
1.3. Wesentliche Erkenntnisse 8

2. Grundlagen und Definitionen 9
2.1. Was verstehen wir unter künstlicher Intelligenz? 9
2.2. Was sind moderne Methoden? 9
2.3. Was verstehen wir unter Automatisierungstools? 10
2.4. Was sind KI-Agenten? 10

3. Regulatorische Anforderungen und ihre Anwendung in der privaten Krankenversicherung 11
3.1. Auswirkungen des AI Acts 11
3.1.1. Klassifizierung von KI-Systemen 11
3.1.2. Zwei relevante Arten von Akteuren: Einsatz als Anbieter oder Betreiber 13
3.1.3. Anforderungen an Hochrisikosysteme 15
3.2. DORA-Governance für Pricing- und KI-Anwendungen im Aktuariat 17
3.3. Einordnung moderner Tools und KI im Zusammenspiel von AI Act und DORA 17
3.4. Anforderungen an Daten und Modelle in der Krankenversicherung 18
3.4.1. Anforderungen im Rahmen der Datennutzung 18
3.4.2. Anforderungen an Modelle und Erklärbarkeit 19

4. Einsatz von künstlicher Intelligenz und modernen Methoden in Kernprozessen der privaten Krankenversicherung 20
4.1. Anwendungsfälle im Kalkulationsablauf der Beitragsanpassung 20
4.1.1. Phase I: Vorbereitung 22
4.1.2. Phase II: Kalkulation – Grundkopfschaden 24
4.1.3. Phase III: Kalkulation – Rechnungsgrundlagen 26
4.1.4. Phase IV: Limitierung und Dokumentation 27
4.1.5. Querschnittsthema: Ablösung von Altsystemen durch eine moderne Kalkulationsumgebung 28
4.2. Anwendungsfälle von modernen Methoden und KI zur Analyse unternehmenseigener Daten und zur Ableitung von Bestandsprognosen 29
4.3. Weitere Anwendungsmöglichkeiten und Ausblick 30

5. Vertiefung ausgewählter Anwendungsfälle 31
5.1. Limitierungstool 31
5.2. Validierung von Technischen Berechnungsgrundlagen 32
5.3. Diagnosedaten in der Schadenbedarfsprognose 34

5.3.1. Ausgangslage ... 34
5.3.2. Der Mehrwert der Aufteilung ... 34
5.3.3. Ein möglicher Ansatz ... 35
5.3.4. Interpretierbare und Black-Box-Modelle ... 35
5.3.5. Herausforderungen und Fazit ... 36
5.4. Fallprognose: Wiedereinweisung ins Krankenhaus ... 36
5.4.1. Schätzung der Wahrscheinlichkeit für eine Wiedereinweisung ... 36
5.4.2. Patientenhistorie als Ausgangspunkt ... 37
5.4.3. Ergebnis: Vergleich von Modellansätzen ... 37
5.5. Analyse der Prämienstruktur eines Krankenversicherungsbestands ... 37
5.5.1. Zielsetzung: Bestandsstrukturen transparent machen ... 37
5.5.2. Datenbasis ... 38
5.5.3. Ergebnisse: Auswirkung struktureller Bestandsmerkmale ... 38

6. Rolle und Verantwortung der Aktuarinnen und Aktuare ... 39
6.1. Einordnung und Bewertung des Einsatzes von KI und modernen Methoden in der privaten Krankenversicherung ... 39
6.2. Verantwortung der Aktuarinnen und Aktuare beim Einsatz von KI ... 39
6.2.1. Einordnung von KI und modernen Methoden und abzuleitende Maßnahmen ... 39
6.2.2. Fachliche Beurteilung wesentlicher Anwendungen ... 41
6.3. Besondere Rolle der Verantwortlichen Aktuarin oder des Verantwortlichen Aktuars beim Einsatz von KI oder modernen Methoden ... 43
6.4. Position der DAV zur Einordnung statistischer Verfahren als KI ... 44

Anhang A. Grundlagen des maschinellen Lernens ... 46
A.1. Einführung ins überwachte maschinelle Lernen ... 46
A.1.1. Generalisierung und Overfitting ... 46
A.1.2. Datenaufbereitung und Visualisierung ... 46
A.1.3. Gütebewertung ... 47
A.2. Prognoseverfahren ... 47
A.2.1. Entscheidungsbäume und Gradient Boosting ... 47
A.2.2. Logistische Regression und Regularisierung ... 48
A.2.3. Künstliche neuronale Netze und Deep Learning ... 48
A.2.4. Blending und Stacking ... 48
A.3. Weitere Themen ... 49
A.3.1. AutoML-Tools ... 49
A.3.2. Erklärbarkeit ... 49
A.3.3. Software und Hardware ... 49

Anhang B. Grundlagen der generativen künstlichen Intelligenz ... 50
B.1. Von Text zu Sprachmodellen ... 50
B.1.1. Tokenisierung und Embeddings ... 50
B.1.2. Vom Zählmodell zum großen Sprachmodell ... 50
B.1.3. Transformer und Attention ... 50
B.2. Steuerung und Anbindung an eigene Daten ... 50
B.2.1. Training, Fine-Tuning und Antwortsteuerung ... 50

B.2.2. Retrieval-Augmented Generation 51
B.2.3. Strukturierte Ausgaben, Werkzeugaufrufe und Agenten 51
B.3. Grenzen und Rahmenbedingungen 51
B.3.1. Halluzinationen, Verzerrungen und Nachvollziehbarkeit 51
B.3.2. Datenschutz, Vertraulichkeit und Sicherheit 51
B.3.3. Kosten, Ressourcen und Einsatzentscheidung 52

Anhang C. Abkürzungsverzeichnis 53

1. Einleitung

Dieses Kapitel umreißt, warum künstliche Intelligenz und moderne Methoden für das Aktuariat der privaten Krankenversicherung zunehmend an Bedeutung gewinnen, welche Ziele der Ergebnisbericht verfolgt und zu welchen wesentlichen Erkenntnissen er gelangt.

1.1. Relevanz der Fragestellung

Künstliche Intelligenz (KI) entwickelt sich in einer Geschwindigkeit, die auch das aktuarielle Arbeiten in der privaten Krankenversicherung spürbar verändert. Während vor wenigen Jahren KI-Tools eher als Zukunftsthema galten, stehen sie heute unmittelbar als Werkzeuge zur Verfügung: zur Datenaufbereitung, zur Modellierung, zur Automatisierung wiederkehrender Aufgaben, zur Dokumentation und sogar zur Unterstützung bei der Programmierung komplexer aktuarieller Tools. Für Aktuarinnen und Aktuare entsteht daraus ein doppelter Handlungsdruck.

Einerseits eröffnet KI erhebliche Chancen: Sie kann die Entwicklung und Kalibrierung von Modellen beschleunigen, erlaubt neue Auswertungs- und Prognoseansätze und unterstützt bei der Bewältigung immer größerer Datenmengen, von Leistungs- und Diagnosedaten über Bestands-, Markt- und externe Gesundheitsdaten bis hin zu Sensitivitätsanalysen im Rahmen der Beitragsanpassung. Wer KI versteht und beherrscht, kann aktuarielles Know-how mit moderner Technologie verbinden und damit fachlich anspruchsvolle Fragestellungen effizienter und umfassender bearbeiten.

Andererseits sind mit dem Einsatz von KI im Aktuariat erhebliche Anforderungen und Risiken verbunden. Die Sicherheit, Verfügbarkeit und Qualität der Daten, die Transparenz der eingesetzten Verfahren, die Verlässlichkeit und Nachvollziehbarkeit von Ergebnissen – all dies sind Kernfragen, die im regulierten Umfeld der Krankenversicherung nicht optional, sondern zwingende Voraussetzungen sind. Hinzu kommen die spezifischen regulatorischen Rahmenbedingungen, etwa im Kontext der Beitragsanpassung, der Verwendung von Gesundheitsdaten oder der Begründung gegenüber Treuhändern und Aufsicht. Black-Box-Verfahren stehen dabei schnell im Spannungsfeld zwischen hoher Prognosegüte und unzureichender Erklärbarkeit. KI-gestützte Programmierung wiederum wirft Fragen nach der verlässlichen Absicherung der fachlichen Logik und gesetzlichen Anforderungen auf.

Gleichzeitig ist der Einsatz von KI nicht völlig ungeregelt. Mit dem Artificial Intelligence Act (AI Act) besteht in der EU ein neuer gesetzlicher und aufsichtsrechtlicher Rahmen, aus dem klare Leitplanken abgeleitet werden können. Zugleich behalten bestehende gesetzliche und aufsichtsrechtliche Anforderungen etwa zu internen Kontrollsystemen oder zu IT-Governance auch bei KI-Einsatz ihre Gültigkeit. Zusammen geben diese Vorgaben Orientierung, wie KI-Tools verantwortungsvoll und regelkonform eingesetzt werden können, ohne auf ihre Vorteile zu verzichten. Sie machen aber auch deutlich, dass Aktuarinnen und Aktuare KI nicht nur anwenden, sondern fachlich und technisch hinreichend verstehen müssen.

Vor diesem Hintergrund hat sich eine Arbeitsgruppe intensiv mit der Rolle von KI und anderen modernen Methoden im Aktuariat der Krankenversicherung beschäftigt.

1.2. Zielsetzung

Der vorliegende Ergebnisbericht soll Leitlinien für einen verantwortungsvollen Einsatz skizzieren, relevante Praxisbeispiele für die Nutzung in Aktuariaten der Krankenversicherung beschreiben und dazu beitragen, dass KI im Aktuariat als fachlich kontrolliertes Werkzeug eingesetzt werden kann.

Dazu werden zunächst die grundlegenden Begriffe und die Abgrenzung zwischen KI- und Automatisierungstools geklärt. Auf dieser Basis erfolgt eine systematische Einordnung der relevanten regulatorischen Anforderungen, insbesondere des AI Acts und des Digital Operational Resilience Acts (DORA), sowie ihre konkrete Anwendung im Umfeld der Krankenversicherung. Hierbei werden Verantwortlichkeiten, insbesondere die Rolle der Verantwortlichen Aktuarinnen und Aktuare,

und die sich aus dem gesetzlichen Rahmen ergebenden Anforderungen an Daten, Modelle und Prozesse herausgearbeitet.

Ein weiterer Schwerpunkt liegt auf der Benennung und Beschreibung konkreter Anwendungsfälle von KI und modernen Methoden in zentralen aktuariellen Kernprozessen, mit Fokus auf den Kalkulationsablauf der Beitragsanpassung sowie die Analyse unternehmenseigener Daten und die Ableitung von Bestandsprognosen. Ausführliche Praxisbeispiele vertiefen diese Anwendungsfelder und illustrieren sowohl den Mehrwert als auch typische Herausforderungen.

Darauf aufbauend beschreibt der Ergebnisbericht die Rolle der Aktuarinnen und Aktuare im KI-Umfeld und leitet Anforderungen an deren Kompetenzen und Arbeitsweise ab. Zwei Anhänge vermitteln die methodischen Grundlagen: Anhang A behandelt das maschinelle Lernen einschließlich typischer Prognoseverfahren, Tools sowie Fragen der Erklärbarkeit und der technischen Umsetzung, Anhang B die Grundlagen der generativen künstlichen Intelligenz.

Auf dieser Grundlage leitet der Ergebnisbericht Leitplanken für die Arbeit der Aktuarinnen und Aktuare ab. Dazu gehören:

- die Identifikation und Bewertung möglicher Methoden und Modellansätze, die sich für den Einsatz im aktuariellen Kontext eignen,
- konkrete Empfehlungen für Auswertungen und Analysen, insbesondere im Hinblick auf Transparenz, Nachvollziehbarkeit und Validierung,
- Hinweise zur angemessenen Dokumentation von KI-gestützten Verfahren und Ergebnissen,
- eine Zusammenstellung der fachlichen, technischen und regulatorischen Anforderungen, die beim Einsatz von KI-Tools erfüllt sein müssen.

1.3. Wesentliche Erkenntnisse

Der Einsatz moderner Datenanalyseverfahren und künstlicher Intelligenz bietet für das Aktuariat in der Krankenversicherung ein großes Potenzial. In nahezu allen Kernprozessen – von der Beitragsanpassung über die Bestands- und Schadenbedarfsprognose bis hin zur Analyse komplexer Datenbestände – können moderne Methoden die Qualität der Ergebnisse erhöhen, Prozesse effizienter gestalten und neue Einblicke in das Geschäftsmodell ermöglichen. Die im Ergebnisbericht dargestellten Anwendungsfälle zeigen, dass sich KI-gestützte Verfahren sowohl zur Unterstützung klassischer aktuarieller Aufgaben als auch zur Erweiterung bestehender Analyse- und Prognoseinstrumente eignen.

Gleichzeitig ist der rechtliche und regulatorische Rahmen anspruchsvoll. AI Act, DORA und die spezifischen Anforderungen der Krankenversicherung stellen Aktuarinnen und Aktuare vor die Aufgabe, neue Tools und Methoden sorgfältig einzuordnen und verantwortungsvoll auszugestalten. Insbesondere die Themen Datenqualität und -schutz, Modellvalidierung, Erklärbarkeit sowie die klare Zuordnung von Verantwortlichkeiten erfordern ein bewusstes und strukturiertes Vorgehen.

Die wesentliche Erkenntnis des Ergebnisberichts ist jedoch, dass diese Herausforderungen beherrschbar sind: Durch gut definierte Prozesse, eine konsequente Dokumentation, klare Governance-Strukturen und gezielte menschliche Kontrollen können Nachvollziehbarkeit und Transparenz in der Regel hinreichend sichergestellt werden. Unter diesen Voraussetzungen lassen sich die mit KI und modernen Tools verbundenen Risiken deutlich reduzieren, sodass ihr Einsatz in einer Vielzahl von Anwendungsfällen nicht nur möglich, sondern fachlich sinnvoll ist.

Für Aktuarinnen und Aktuare bedeutet dies: KI und moderne Methoden sind leistungsfähige Werkzeuge, die – eingebettet in einen rechtssicheren und gut dokumentierten Rahmen – einen messbaren Mehrwert für Kalkulation, Analyse und Steuerung in der Krankenversicherung liefern können. Der Ergebnisbericht zeigt auf, wo diese Potenziale liegen, welche Anforderungen zu erfüllen sind und wie die Praxis so gestaltet werden kann, dass Innovation und Verantwortung miteinander vereinbar bleiben.

2. Grundlagen und Definitionen

Dieses Kapitel klärt die zentralen Begriffe des Ergebnisberichts – künstliche Intelligenz, moderne Methoden, Automatisierungstools und KI-Agenten – und grenzt sie voneinander ab.

2.1. Was verstehen wir unter künstlicher Intelligenz?

In der Literatur finden sich zahlreiche Definitionen von künstlicher Intelligenz. Da die Krankenversicherung letztlich ein Rechtsgeschäft darstellt, ist an dieser Stelle besonders von Bedeutung, wie künstliche Intelligenz in rechtlicher Hinsicht definiert ist. Die Europäische Union (EU) hat mit dem Artificial Intelligence Act (AI Act) einen unionsweit verbindlichen Rechtsrahmen für die Entwicklung und den Einsatz von künstlicher Intelligenz geschaffen. Der AI Act liefert damit zentrale Vorgaben und ist die wesentliche Grundlage der weiteren Betrachtungen.

In Artikel 3 Nummer 1 des AI Acts heißt es:

> *„KI-System“: ein maschinengestütztes System, das für einen in unterschiedlichem Grade autonomen Betrieb ausgelegt ist und das nach seiner Betriebsaufnahme anpassungsfähig sein kann und das aus den erhaltenen Eingaben für explizite oder implizite Ziele ableitet, wie Ausgaben wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erstellt werden, die physische oder virtuelle Umgebungen beeinflussen können;*

Die DAV hat sich eingehend mit den Auswirkungen des AI Acts auf die Versicherungsbranche und die Arbeit von Aktuarinnen und Aktuaren beschäftigt. Auf den Ergebnisbericht „Der Artificial Intelligence Act im aktuariellen Kontext“ sei verwiesen.[5] Den dortigen Ausführungen liegt ebenfalls die oben angegebene Definition zugrunde. Ergänzend liegt der Chancen- und Risikobericht des Ausschusses Enterprise Risk Management vor, der die Auswirkungen des AI Acts im Hinblick auf Chancen, Risiken und insbesondere Emerging Risks vertieft.[6]

In der Praxis lässt sich zwischen herkömmlichen EDV-Algorithmen und KI-Anwendungen (d. h. Anwendungen, die auf KI-Systemen beruhen) bislang keine trennscharfe Grenze ziehen. Es ist daher nicht immer eindeutig, auf welche EDV-Anwendungen die aus dem AI Act abzuleitenden Anforderungen anzuwenden sind.

Wo im Folgenden von „maschinellem Lernen“ die Rede ist, zählen auch diese Anwendungen zur „künstlichen Intelligenz“ im oben genannten Sinne; dies gilt ebenso für große Sprachmodelle.

2.2. Was sind moderne Methoden?

Moderne Methoden bezeichnen den Einsatz fortschrittlicher Technologien und Werkzeuge, um fachliche Aufgaben effizienter umzusetzen und gleichzeitig konkrete Anforderungen einfacher und schneller in praktikable Lösungen zu überführen. Dazu gehört auch, dass fachliche Anliegen gezielt unterstützt und automatisiert umgesetzt werden können. Diese können sowohl mit als auch ohne KI-Unterstützung aufgebaut sein.

Beispielsweise lassen sich im Kontext einer Beitragsanpassung solche KI-gestützte, moderne Methoden vielfältig einsetzen. In Abschnitt 4.1 wird der sinnvolle Einsatz moderner Methoden und KI entlang eines in vier Phasen gegliederten Kalkulationsablaufs systematisch dargestellt. Dabei werden verschiedene Verfahren genannt, etwa Machine-Learning-Verfahren wie Generalisierte Additive Modelle (GAMs); eine kurze Einführung in diese Methoden gibt Anhang A. Solche Verfahren eignen sich besonders, um komplexe Datenstrukturen zu analysieren und künftige Entwicklungen genauer zu prognostizieren. KI kann dabei entlang des gesamten Prozesses unterstützen, indem sie Daten aufbereitet, Berechnungen vorbereitet, Zwischenergebnisse verständlich darstellt oder Dokumentationsentwürfe erstellt.

[5] Siehe Ergebnisbericht „Der Artificial Intelligence Act im aktuariellen Kontext“ des Ausschusses ADS / AI in der Fassung vom 11.03.2025.

[6] Siehe Ergebnisbericht „Ergänzender Chancen- und Risikobericht zu dem AI Act“ des Ausschusses Enterprise Risk Management in der Fassung vom 19.05.2025.

Ein zusätzlicher zentraler Baustein moderner KI-gestützter Methoden sind sogenannte große Sprachmodelle („Large Language Models“, LLMs). Sie ermöglichen es, mit der KI über Eingaben in natürlicher Sprache zu interagieren. So können Nutzerinnen und Nutzer fachliche Anforderungen, wie etwa Berechnungslogiken, Modelle oder gewünschte Visualisierungen, in natürlicher Sprache beschreiben. Das LLM verarbeitet die Beschreibung und setzt sie etwa in Programmcode um. Dadurch wird es möglich, ohne tiefgehende Programmierkenntnisse schnell erste Anwendungen, Auswertungen oder Visualisierungen zu erstellen. Ein mögliches Anwendungsfeld wird in Abschnitt 5.1 „Limitierungstool“ aufgezeigt.

Ein weiteres Beispiel ist eine LLM-gestützte Validierung von Technischen Berechnungsgrundlagen mit einer automatischen Konsistenzprüfung, bei der Inhalte auf Widersprüche, Vollständigkeit und logische Plausibilität überprüft werden. Hierauf wird in Abschnitt 5.2 „Validierung von Technischen Berechnungsgrundlagen“ näher eingegangen.

Somit ist es möglich, eine direkte Verbindung zwischen fachlicher Idee und technischer Realisierung herzustellen – das beschleunigt die Umsetzung erheblich. Gleichzeitig bleibt die Verantwortung beim Menschen: Die Aktuarin bzw. der Aktuar definiert die Anforderungen, prüft die Ergebnisse und stellt die fachliche Korrektheit sicher. Die KI dient als Werkzeug, das die technische Umsetzung unterstützt und den Entwicklungsprozess deutlich effizienter gestaltet.

2.3. Was verstehen wir unter Automatisierungstools?

Unter Automatisierungstools verstehen wir IT-Programme, die wiederkehrende und regelbasierte Aufgaben effizient ausführen. Ziel ist es, manuelle Tätigkeiten zu reduzieren, Abläufe zu standardisieren und die Reproduzierbarkeit von Ergebnissen zu erhöhen. Typische Beispiele sind automatisierte Datenaufbereitung, standardisierte Berechnungsschritte, wiederkehrende Reportings oder die Erstellung von Dokumentationen nach festen Regeln. Wir verstehen derartige Automatisierungstools als Teilmenge der modernen Methoden.

Charakteristisch für Automatisierungstools ist ihre deterministische Funktionsweise: Es werden klare Regeln und Rechenschritte vorgegeben, sodass sie bei gleichen Eingaben immer exakt das gleiche Ergebnis liefern. Voraussetzung solcher Tools sind stabile und gut strukturierte Prozesse. Im Gegensatz dazu zielt der Einsatz von KI nicht primär auf die reine Automatisierung ab, sondern auf die Erweiterung und Verbesserung von Entscheidungen und Analysen. KI-Modelle sind häufig probabilistisch und datengetrieben, was eine Nachvollziehbarkeit und Reproduzierbarkeit erschwert. Durch den Einsatz von KI können beispielsweise Muster in großen Datenmengen und Zusammenhänge erkannt werden, ohne explizite Vorgabe von Regeln.

Automatisierungstools und KI-Methoden schließen sich nicht gegenseitig aus: Ein sinnvoller Einsatz von KI-Methoden erfolgt oftmals eingebettet in bestehende automatisierte Lösungen – etwa bei der in Abschnitt 5.2 beschriebenen Validierung Technischer Berechnungsgrundlagen, die regelbasierte Prüfungen mit LLM-gestützten Prüfschritten verbindet.

2.4. Was sind KI-Agenten?

KI-Agenten sind Systeme, die KI-Modelle, in der Regel LLMs, mit definierten Werkzeugen kombinieren. Sie planen mehrstufige Aufgaben zielgerichtet, rufen je nach Kontext passende Tools auf, interagieren mit Dokumenten und Systemen und können Ergebnisse iterativ verbessern. Sie stellen damit eine spezielle Form von KI-Anwendungen dar.

Im Unterschied zu klassischen Automatisierungslösungen sind Agenten nicht ausschließlich regelbasiert und deterministisch, sondern arbeiten probabilistisch und kontextsensitiv. Dadurch sind sie bei unstrukturierten Informationen, variierenden Eingaben oder Ausnahmen flexibler als klassische Automatisierungstools und erweisen sich zunehmend als performante Lösung in komplexen Wissens- und Analyseprozessen.

Eine Herausforderung stellt die eingeschränkte Kontrollierbarkeit dar. Richtig eingebettet in Governance, Datenschutz und Modellrisikomanagement bieten sie dennoch für aktuarielle Arbeitsprozesse ein hohes Maß an Effizienzpotenzial.

3. Regulatorische Anforderungen und ihre Anwendung in der privaten Krankenversicherung

Dieses Kapitel ordnet die regulatorischen Anforderungen an KI und moderne Methoden ein – insbesondere den AI Act und DORA – und leitet daraus konkrete Anforderungen an Daten, Modelle und Prozesse ab.

Der für aktuarielle Aspekte in der Krankenversicherung durch VVG, VAG, KVAV gegebene Rechtsrahmen wird als bekannt vorausgesetzt und deshalb an dieser Stelle nicht weiter ausgeführt.

3.1. Auswirkungen des AI Acts

Wird bei aktuariellen Tätigkeiten künstliche Intelligenz entwickelt oder angewendet, gelten besondere Vorschriften. Der Rat der Europäischen Union hat die Verordnung (EU) 2024/1689 vom 13. Juni 2024 (AI Act, KI-Verordnung) am 21. Mai 2024 gebilligt. Sie trat am 1. August 2024 in Kraft. Diese Verordnung ist in allen ihren Teilen verbindlich und gilt unmittelbar in jedem Mitgliedstaat. Übergangsregelungen für bestimmte Bereiche sind vorgesehen.

Die wesentlichen Zielsetzungen des AI Acts ergeben sich insbesondere aus Artikel 1 sowie den Erwägungsgründen 1, 6 bis 8 und 27 der Verordnung:

- Der AI Act soll die Einführung menschenzentrierter und vertrauenswürdiger KI fördern und zugleich ein hohes Schutzniveau für Gesundheit, Sicherheit und Grundrechte gewährleisten. Durch einheitliche Anforderungen soll das Vertrauen in KI-Systeme gestärkt und deren verantwortungsvolle Entwicklung und Verwendung unterstützt werden.
- KI-Systeme könnten bei unkontrollierter Anwendung die menschlichen Grundrechte verletzen. Durch den AI Act soll diese Gefahr ausgeschlossen oder zumindest verringert werden.
- Die hohe Komplexität von KI-Systemen kann zu intransparenten Ergebnissen führen. Durch den AI Act soll sichergestellt werden, dass über die Ergebnisse Rechenschaft abgelegt werden kann.
- Der AI Act soll den durch KI ausgelösten Wandel der Arbeitswelt sachgerecht begleiten und dabei Risiken für Beschäftigte wie wirtschaftliche Chancen berücksichtigen.

Eine vertiefte spartenübergreifende Darstellung der Auswirkungen des AI Acts auf die aktuarielle Tätigkeit enthalten die Ergebnisberichte der Ausschüsse ADS / AI und Enterprise Risk Management.[7][8] Im Folgenden werden die für die private Krankenversicherung besonders relevanten Anforderungen zusammengefasst.

3.1.1. Klassifizierung von KI-Systemen

Der AI Act sieht eine vierstufige Klassifizierung von KI-Systemen nach deren Risikopotenzial vor (siehe Abbildung 1). Demnach gibt es:

- verbotene Praktiken,
- Hochrisikosysteme,
- KI-Systeme mit besonderen Transparenzanforderungen,
- alle weiteren KI-Systeme.

[7] Siehe Ergebnisbericht „Der Artificial Intelligence Act im aktuariellen Kontext“ des Ausschusses ADS / AI in der Fassung vom 11.03.2025.

[8] Siehe Ergebnisbericht „Ergänzender Chancen- und Risikobericht zu dem AI Act“ des Ausschusses Enterprise Risk Management in der Fassung vom 19.05.2025, abrufbar unter https://aktuar.de/de/wissen/fachinformationen/detail/ergaenzender-chancen-und-risikobericht-zu-dem-ai-act/.

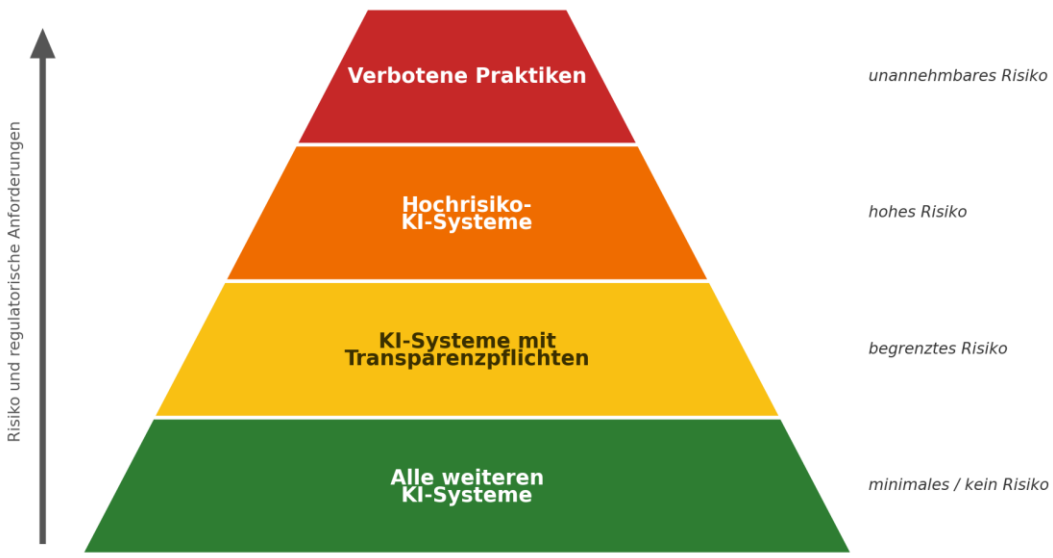


Abbildung 1: Risikobasierte Klassifizierung von KI-Systemen nach dem AI Act (vereinfachte Darstellung).

Die Zuordnung eines KI-Systems zu einer dieser Klassen erfolgt anhand von festgelegten Merkmalen bzw. Einsatzzwecken.

Gemäß Anhang III des AI Acts gelten insbesondere KI-Systeme aus dem Bereich „Zugang zu und Inanspruchnahme von wesentlichen privaten Dienstleistungen und wesentlichen öffentlichen Diensten und Leistungen“ als Hochrisikosysteme. Unter Anhang III Nummer 5 Buchstabe c werden explizit KI-Systeme genannt, „die bestimmungsgemäß für die Risikobewertung und Preisbildung in Bezug auf natürliche Personen im Fall von Lebens- und Krankenversicherungen verwendet werden sollen“. Sofern keine verbotene Praxis nach Artikel 5 des AI Acts vorliegt, fallen KI-Systeme, die bestimmungsgemäß der Risikobewertung oder Preisbildung in Bezug auf natürliche Personen in der Krankenversicherung dienen, damit grundsätzlich in den Anwendungsbereich der Hochrisikoklassifizierung; mögliche Ausnahmen nach Artikel 6 Absatz 3 sind im Einzelfall zu prüfen.

In Artikel 6 Absatz 3 des AI Acts werden jedoch Ausnahmen für diese Zuordnung genannt. Ein System gilt dann nicht als hochriskant, „wenn es kein erhebliches Risiko der Beeinträchtigung in Bezug auf die Gesundheit, Sicherheit oder Grundrechte natürlicher Personen birgt, indem es unter anderem nicht das Ergebnis der Entscheidungsfindung wesentlich beeinflusst“. Dies ist insbesondere dann der Fall, wenn z. B. das KI-System dazu bestimmt ist,

- das Ergebnis einer zuvor abgeschlossenen menschlichen Tätigkeit zu verbessern oder
- lediglich eine vorbereitende Aufgabe für eine Bewertung durchzuführen, die für die Zwecke der in Anhang III des AI Acts aufgeführten Anwendungsfälle relevant ist.

In diesen Fällen schreibt Artikel 6 Absatz 4 des AI Acts eine Dokumentation der Bewertung und die Registrierung des KI-Systems vor.

Auch wenn ein KI-System lediglich unterstützend bei einer menschlichen Tätigkeit eingesetzt wird, gilt es dennoch als hochriskant, wenn es ein Profiling natürlicher Personen vornimmt. Unter Profiling fällt jede Art der automatisierten Verarbeitung personenbezogener Daten, die diese verwendet, um persönliche Aspekte wie z. B. Gesundheitsdaten zu analysieren, zu bewerten oder vorherzusagen. Unter diesen Gegebenheiten sind auch diejenigen KI-Systeme, die nur unterstützend im Zuge einer Risikoprüfung bei Abschluss eines Krankenversicherungsvertrages eingesetzt werden, trotz der genannten Ausnahmeregelung stets als hochriskant einzustufen. Zu den daraus folgenden praktischen Anforderungen an den Umgang mit (Gesundheits-)Daten siehe Abschnitt 3.4.1.

3.1.2. Zwei relevante Arten von Akteuren: Einsatz als Anbieter oder Betreiber

Der AI Act unterscheidet verschiedene Akteure (neben Anbietern und Betreibern etwa Einführer, Händler und Bevollmächtigte), denen abhängig von ihrer Rolle unterschiedliche Pflichten zugewiesen werden. Für aktuarielle Anwendungen in Versicherungsunternehmen sind vor allem Anbieter und Betreiber relevant.

Anbieter ist, wer ein KI-System entwickelt oder entwickeln lässt und es unter eigenem Namen oder eigener Handelsmarke in Verkehr bringt oder in Betrieb nimmt. Betreiber ist, wer ein KI-System in eigener Verantwortung im Rahmen seiner beruflichen Tätigkeit verwendet.

Diese Rollen liegen regelmäßig beim Versicherungsunternehmen als juristischer Person. Die einzelne Aktuarin bzw. der einzelne Aktuar ist daher üblicherweise weder Anbieter noch Betreiber im Rechtssinn. Sie oder er kann innerhalb des Unternehmens jedoch Aufgaben wahrnehmen, die der Anbieter- oder Betreiberseite zuzuordnen sind. Hierzu gehören etwa die Festlegung fachlicher Anforderungen, die Entwicklung und Validierung eines Modells, die Prüfung seiner Ausgaben oder die Überwachung im laufenden Betrieb.

Unabhängig von der Risikoklasse verpflichtet Artikel 4 des AI Acts Anbieter und Betreiber, für ein ausreichendes Maß an KI-Kompetenz des mit KI befassten Personals zu sorgen – also für die Fähigkeiten, Kenntnisse und das Verständnis, um KI-Systeme informiert einzusetzen und ihre Chancen, Grenzen und Risiken einzuschätzen. Für Aktuarinnen und Aktuare, die KI-gestützte Verfahren entwickeln oder anwenden, ist der Aufbau dieser Kompetenz damit auch eine regulatorische Anforderung; die Pflicht gilt seit dem 2. Februar 2025.

Abgrenzung der Rollen

Ein Versicherungsunternehmen handelt regelmäßig als Betreiber, wenn es ein von einem Dritten bereitgestelltes KI-System verwendet. Beispiele aus der aktuariellen Praxis sind:

- die Nutzung eines externen Sprachmodells zur Erstellung von Text- oder Programmentwürfen
- der Einsatz einer eingekauften Anwendung zur Erkennung auffälliger Schadendaten
- die Nutzung eines fremden Prognosemodells
- die Verwendung eines KI-basierten Programmierassistenten

Eine Anbieterrolle kommt in Betracht, wenn das Versicherungsunternehmen ein KI-System selbst entwickelt oder entwickeln lässt. Dies kann beispielsweise der Fall sein bei:

- einem im Aktuariat entwickelten Prognosemodell
- einer eigenen Anwendung zur Erkennung auffälliger Schäden
- einem unternehmensintern entwickelten System zur Prüfung Technischer Berechnungsgrundlagen
- einer Fachanwendung, die ein externes Sprachmodell mit unternehmenseigenen Datenquellen, Prüfregeln und Verarbeitungsschritten verbindet.

Für die Anbieterrolle ist weder erforderlich, dass das Versicherungsunternehmen alle verwendeten Komponenten selbst entwickelt, noch dass es die Anwendung Dritten anbietet. Auch eine ausschließlich unternehmensintern eingesetzte Anwendung kann eine Anbieterrolle begründen, wenn das Versicherungsunternehmen das KI-System entwickelt oder entwickeln lässt und unter eigenem Namen oder eigener Marke für den Eigengebrauch in Betrieb nimmt. Es kann daher Anbieter einer eigenen Fachanwendung sein, obwohl diese ein Modell oder einen Dienst eines externen Anbieters nutzt. Der externe Dienstleister bleibt Anbieter des von ihm bereitgestellten Modells oder Systems, während das Versicherungsunternehmen Anbieter des darauf aufbauenden eigenen KI-Systems sein kann.

Unabhängig davon kann ein Versicherungsunternehmen nach Artikel 25 Absatz 1 des AI Acts auch nachträglich zum Anbieter eines bereits bestehenden KI-Systems werden. Dies ist insbesondere dann der Fall, wenn

- es eine wesentliche Veränderung eines Hochrisiko-KI-Systems, das bereits in Verkehr gebracht oder in Betrieb genommen wurde, so vornimmt, dass es weiterhin ein Hochrisiko-KI-System bleibt, oder
- die Zweckbestimmung eines KI-Systems, einschließlich eines KI-Systems mit allgemeinem Verwendungszweck, das nicht als hochriskant eingestuft wurde, so verändert, dass das betreffende KI-System zu einem Hochrisiko-KI-System wird.

Bei internen Eigenentwicklungen kann das Versicherungsunternehmen beide Rollen zugleich einnehmen. Es ist Anbieter, weil es das System entwickelt oder entwickeln lässt und wird durch die anschließende Verwendung zugleich Betreiber.

Für Aktuarinnen und Aktuare ist die Abgrenzung der konkreten Rolle insbesondere dann relevant, wenn sie ein Modell nicht nur anwenden, sondern dessen fachlichen Verwendungszweck, Datenbasis und Funktionsweise maßgeblich bestimmen. Dabei sind vor allem folgende Fragen zu betrachten:

- Wer legt fest, welche fachliche Aufgabe das System erfüllen soll?
- Wer bestimmt die fachlichen und technischen Anforderungen?
- Wer entscheidet über Modell, Merkmale und Daten?
- Wer legt die Anforderungen an Genauigkeit und Validierung fest?
- Wer entscheidet über die Freigabe und Inbetriebnahme?
- Unter wessen Namen wird das System eingesetzt?
- Wer entscheidet über wesentliche Änderungen?
- Wer entscheidet über die Aussetzung der Verwendung, die Abschaltung oder die endgültige Stilllegung des Systems?

Keines dieser Kriterien entscheidet die Rollenfrage isoliert. Maßgeblich ist eine Gesamtbetrachtung. Die Bereitstellung von Daten, die Formulierung einzelner Anforderungen oder die Einstellung vorgesehener Parameter führt für sich genommen noch nicht zur Anbieterstellung. Je stärker das Versicherungsunternehmen Zweck, Entwicklung und Funktionsweise des Systems prägt, desto eher liegt eine Anbieterrolle vor.

Allgemeine Sprachmodelle

Bei allgemeinen Sprachmodellen ist die Abgrenzung weniger eindeutig als bei spezialisierten Fachanwendungen. Solche Modelle sind für eine Vielzahl von Aufgaben ausgelegt und verfügen regelmäßig über keinen einzelnen fachlichen Verwendungszweck. Der AI Act berücksichtigt sie durch ein eigenes Regelungsregime für Modelle mit allgemeinem Verwendungszweck, das sich vor allem an deren Anbieter richtet.

Stellt ein Versicherungsunternehmen seinen Aktuarinnen und Aktuaren einen externen, auf einem allgemeinen Sprachmodell beruhenden KI-Dienst als Arbeitsmittel zur Verfügung, ist es hinsichtlich der Verwendung dieses Dienstes grundsätzlich Betreiber. Dies gilt beispielsweise für die sprachliche Überarbeitung eines Textes, die Unterstützung bei der Programmierung oder die Strukturierung einer fachlichen Fragestellung.

Die allgemeine Freigabe eines solchen KI-Dienstes besagt allerdings noch nicht, dass jeder aktuarielle Anwendungsfall geprüft wurde. Die Verwendung für eine Kalkulation oder für die Herleitung von Rechnungsgrundlagen kann andere Anforderungen auslösen als die Erstellung eines Textentwurfs. Der konkrete Einsatz ist deshalb gesondert einzuordnen.

Bindet das Unternehmen ein allgemeines Sprachmodell in eine eigene Fachanwendung ein, ist davon auszugehen, dass sich die Rollenbeurteilung ändert. Verarbeitet die Anwendung beispielsweise selbstständig Kalkulationsdaten, ruft sie ein Sprachmodell auf, prüft dessen Ausgaben anhand eigener Regeln und erzeugt daraus einen strukturierten Prüfbericht, kann das Versicherungsunternehmen Anbieter des Gesamtsystems sein.

Bei allgemeinen Sprachmodellen bestehen die Informationen für den Nutzer häufig aus Produktdokumentationen, Nutzungsbedingungen und technischen Hinweisen. Eine fachbezogene Betriebsanleitung für aktuarielle Anwendungen liegt üblicherweise nicht vor. Die besondere Pflicht zur Verwendung entsprechend einer Betriebsanleitung betrifft vor allem Betreiber von Hochrisiko-KI-Systemen, auf die im folgenden Abschnitt näher eingegangen wird. Sie lässt sich nicht pauschal auf jede Nutzung eines allgemeinen Sprachmodells übertragen.

3.1.3. Anforderungen an Hochrisikosysteme

Für Hochrisiko-KI-Systeme gelten umfassende Anforderungen an Gestaltung, Entwicklung und Verwendung. Ein Teil ist bereits vor der Inbetriebnahme zu erfüllen und betrifft vor allem den Anbieter. Weitere Pflichten knüpfen an die konkrete Verwendung an und richten sich an den Betreiber.

Anforderungen an das System und Pflichten des Anbieters

Die Artikel 9 bis 15 AI Act legen die grundlegenden Anforderungen an Hochrisiko-KI-Systeme fest:

- **Risikomanagement:** Für das System ist ein fortlaufendes Risikomanagement einzurichten. Es umfasst die Ermittlung und Bewertung vorhersehbarer Risiken, geeignete Gegenmaßnahmen sowie Tests während der Entwicklung und vor der Inbetriebnahme. Bei aktuariellen Anwendungen sind unter anderem ungeeignete Daten, instabile Ergebnisse, eine unzureichende Übertragbarkeit auf Teilbestände und mögliche Auswirkungen auf Risikobewertung oder Preisbildung zu betrachten.

- **Datenqualität und Daten-Governance:** Trainings-, Validierungs- und Testdaten müssen für den vorgesehenen Zweck relevant und hinreichend repräsentativ sein. Sie sollen so weit wie möglich vollständig und fehlerfrei sein. Herkunft, Aufbereitung, Speicherung und mögliche Verzerrungen sind zu berücksichtigen. In der Krankenversicherung treten die besonderen datenschutzrechtlichen Anforderungen an Gesundheitsdaten hinzu. Die Bereitstellung von Trainingsdaten durch ein Versicherungsunternehmen führt allein noch nicht zur Anbieterstellung. Maßgeblich bleibt, wer das System entwickelt oder entwickeln lässt und unter eigenem Namen in Betrieb nimmt.

- **Dokumentation und Protokollierung:** Vor der Inbetriebnahme ist eine technische Dokumentation zu erstellen und aktuell zu halten. Sie muss insbesondere Verwendungszweck, Systemarchitektur, Daten, Entwicklungsmethoden, Leistungskennzahlen, Validierung, Risikomanagement und Maßnahmen der menschlichen Aufsicht beschreiben. Das System muss außerdem über Funktionen zur automatischen Aufzeichnung relevanter Ereignisse verfügen.

- **Transparenz und Betriebsanleitung:** Das System muss so gestaltet sein, dass der Betreiber seine Ausgaben angemessen interpretieren und verwenden kann. Die Betriebsanleitung muss unter anderem den Verwendungszweck, die getestete Genauigkeit, bekannte Grenzen, Anforderungen an Eingabedaten, vorhersehbare Fehlanwendungen und Maßnahmen der menschlichen Aufsicht erläutern.
- **Menschliche Aufsicht:** Das System muss eine wirksame Kontrolle durch natürliche Personen ermöglichen. Die Aufsichtspersonen müssen die Fähigkeiten und Grenzen des Systems verstehen, Ausgaben sachgerecht interpretieren und Ergebnisse zurückweisen oder übersteuern können. Bei Bedarf müssen sie die Verwendung aussetzen oder das System anhalten können. Dabei ist auch der Gefahr Rechnung zu tragen, dass Nutzer maschinellen Ergebnissen ein unangemessen hohes Vertrauen entgegenbringen.
- **Genauigkeit, Robustheit und Cybersicherheit:** Das System muss während seines Lebenszyklus ein für den Verwendungszweck angemessenes Maß an Genauigkeit, Robustheit und Cybersicherheit erreichen. Die getestete Genauigkeit und die verwendeten Kennzahlen sind in der Betriebsanleitung anzugeben. Für aktuarielle Anwendungen können Prognosefehler, Trefferquoten, Stabilitätskennzahlen und Auswirkungen auf Prämien oder Teilbestände relevant sein.

Zusätzlich muss der Anbieter ein Qualitätsmanagementsystem einrichten, die Konformitätsbewertung durchführen, die erforderliche Dokumentation aufbewahren und das System nach seiner Inbetriebnahme beobachten. Hinzu kommen Pflichten zur Registrierung, zu Korrekturmaßnahmen und zur Meldung schwerwiegender Vorfälle.

Im Übrigen ist eine Einstufung als Nicht-Hochrisiko-KI-System durch den Anbieter zu dokumentieren.

Pflichten des Betreibers

Betreiber müssen Hochrisiko-KI-Systeme entsprechend der Betriebsanleitung verwenden. Sie übertragen die menschliche Aufsicht natürlichen Personen, die über die erforderliche Kompetenz, Ausbildung und Befugnis sowie über die erforderliche Unterstützung verfügen. Die erforderliche Unterstützung umfasst insbesondere angemessene zeitliche, technische und organisatorische Ressourcen, den Zugang zu den benötigten Informationen sowie geeignete Eingriffs- und Eskalationsmöglichkeiten. Die nach Artikel 4 des AI Acts erforderliche KI-Kompetenz ist dabei sicherzustellen. Soweit das Versicherungsunternehmen die Eingabedaten kontrolliert, muss es deren Relevanz und hinreichende Repräsentativität für den Verwendungszweck sicherstellen. Dies ist für Aktuarinnen und Aktuare insbesondere bedeutsam, wenn ein eingekauftes System auf eigene Bestands-, Schaden- oder Diagnosedaten angewendet wird.

Der Betrieb ist anhand der Betriebsanleitung zu überwachen. Werden Risiken oder schwerwiegende Vorfälle erkannt, muss das Unternehmen die vorgesehenen Informations- und Eskalationswege einhalten und die Verwendung erforderlichenfalls aussetzen. Die automatisch erzeugten Protokolle sind, soweit sie unter der Kontrolle des Betreibers stehen, grundsätzlich mindestens sechs Monate aufzubewahren.

Das Versicherungsunternehmen hat als Betreiber eines Hochrisiko-KI-Systems zur Risikobewertung oder Preisbildung in der Krankenversicherung vor dessen erster Verwendung eine Grundrechte-Folgenabschätzung nach Artikel 27 des AI Acts durchzuführen. Sie umfasst insbesondere den Einsatzprozess, die betroffenen Personengruppen, mögliche Beeinträchtigungen, die menschliche Aufsicht und die vorgesehenen Abhilfemaßnahmen. Die Beurteilung ist bei wesentlichen Änderungen zu aktualisieren.

Die Pflichten des Betreibers sind also gegenüber denen des Anbieters eingeschränkt, umfassen jedoch weiterhin wesentliche Punkte, die über das Maß beim Einsatz von klassischen Anwendungen der Informations- und Kommunikationstechnologie (IKT) hinausgehen.

3.2. DORA-Governance für Pricing- und KI-Anwendungen im Aktuariat

Mit der Verordnung (EU) 2022/2554 („Digital Operational Resilience Act“, DORA) gilt seit dem 17. Januar 2025 ein EU-weit einheitlicher Rahmen zur digitalen operationalen Resilienz von Finanzunternehmen. Für Versicherer ersetzt DORA den bisherigen aufsichtsrechtlichen Rahmen der versicherungsaufsichtlichen Anforderungen an die IT (VAIT; bislang Bafin-Verwaltungspraxis in Form eines Rundschreibens, das die Bafin mit Geltungsbeginn von DORA aufgehoben hat) und bildet damit den maßgeblichen Bezugspunkt für die Steuerung von IKT-Risiken. Ziel ist es, IT-gestützte Prozesse auch bei Störungen, technischen Fehlern, Cyberangriffen oder Abhängigkeiten von Dienstleistern beherrschbar, überwachbar und wiederherstellbar zu halten.

Für Pricing-, Kalkulations- und KI-Anwendungen im Aktuariat bedeutet dies, dass neben dem mathematischen Modell auch der gesamte technische und organisatorische Rahmen angemessen zu steuern ist. Wesentliche Anforderungen betreffen insbesondere:

- klare Verantwortlichkeiten für fachlichen Betrieb, technische Umsetzung, Änderungen und Freigaben,
- kontrollierte Änderungsprozesse einschließlich dokumentierter Gründe, Tests, Freigaben und Rückfallmöglichkeiten,
- Nachvollziehbarkeit und Dokumentation von Datenständen, Modell- bzw. Programmversionen, Parametern und Ergebnissen,
- Störungs- und Incident-Prozesse einschließlich Ursachenanalyse, Maßnahmenverfolgung und ggf. Meldepflichten,
- Notfallvorsorge und Wiederanlauf kritischer Anwendungen,
- Steuerung externer Abhängigkeiten, etwa bei Cloud-Plattformen, externen Tools, Datenanbietern oder ausgelagerten Komponenten.

Nach dem Verständnis der Bafin sind KI-Systeme regelmäßig als IKT-Systeme im Sinne von DORA einzuordnen. Damit unterliegen auch KI-gestützte Anwendungen einer risikobasierten Governance über den gesamten Lebenszyklus – von Konzeption, Datenbeschaffung und Entwicklung über Betrieb, Monitoring und Änderungen bis zur Stilllegung. Die Kontrolltiefe richtet sich nach Kritikalität und Ergebniswirkung der Anwendung. Eine kritische Excel-/VBA-Anwendung kann daher ähnlichen Dokumentations-, Test- und Freigabeanforderungen unterliegen wie ein modernes Machine-Learning-Modell. Moderne Plattformen können die Nachweisführung allerdings durch Versionierung, automatisierte Tests, Protokollierung und reproduzierbare Läufe erleichtern.

Für das Aktuariat folgt daraus eine praxisnahe Leitlinie: Pricing- und KI-Anwendungen sollten als fachlich-kritische IKT-Anwendungen behandelt werden, sofern sie Auswirkungen auf Prämien, Rückstellungen, Limitierung oder aufsichtsrelevante Nachweise haben. Die Ausgestaltung der Kontrollen ist risikobasiert vorzunehmen, wobei fachliche Richtigkeit, technische Stabilität und Nachvollziehbarkeit gemeinsam abgesichert werden sollten.

3.3. Einordnung moderner Tools und KI im Zusammenspiel von AI Act und DORA

DORA und AI Act verfolgen unterschiedliche Schutzziele, ergänzen sich in der praktischen Umsetzung aber wesentlich. DORA adressiert die digitale operationale Resilienz, also insbesondere Verfügbarkeit, Sicherheit, Beherrschbarkeit und Wiederherstellbarkeit IKT-gestützter Prozesse. Der AI Act fokussiert demgegenüber Risiken von KI-Systemen, insbesondere bei Hochrisiko-KI, und stellt Anforderungen an Datenqualität, Transparenz, menschliche Aufsicht, Robustheit und Dokumentation. Für KI-gestützte Kalkulationsverfahren im KV-Aktuariat ergibt sich daraus ein integrierter Governance-Ansatz, der sowohl die betriebliche Resilienz als auch KI-spezifische Qualitäts- und Aufsichtsanforderungen abdeckt.

Bei der regulatorischen Einordnung ist zunächst zu unterscheiden, ob im produktiven Betrieb tatsächlich ein KI-System im Sinne des AI Acts eingesetzt wird. Wird beispielsweise generative KI

lediglich zur Erstellung von Quellcode genutzt, das produktive Programm selbst arbeitet aber deterministisch nach festen Regeln, ist dieses produktive Programm regelmäßig kein KI-System. In diesem Fall stehen nicht die Risikoklassen des AI Acts, sondern die DORA-Anforderungen an Governance, Tests, Dokumentation, Betrieb und Änderungsmanagement im Vordergrund. Gleichwohl kann die KI-gestützte Codeerstellung neue Risiken begründen, etwa dadurch, dass generierter Code auf den ersten Blick korrekt wirkt, die fachliche Logik aber fehlerhaft umsetzt. Daher gewinnen fachliche Spezifikation, unabhängige Validierung und reproduzierbare Testfälle besonders an Bedeutung.

Sowohl DORA als auch der AI Act sind risikobasiert anzuwenden. Unter DORA richtet sich die Kontrolltiefe nach Bedeutung und Kritikalität der Anwendung, unter dem AI Act insbesondere danach, ob ein KI-System als Hochrisiko-KI einzustufen ist oder eine Ausnahme greift. Praktisch sollte deshalb je Anwendung nachvollziehbar dokumentiert werden:

- ob ein KI-System im Sinne des AI Acts vorliegt,
- ob eine Hochrisiko-Einstufung oder eine begründete Ausnahme einschlägig ist,
- welche DORA-Kritikalität die Anwendung besitzt,
- welche fachlichen, technischen und organisatorischen Kontrollen daraus folgen.

Ein zweckmäßiges Zielbild ist ein gemeinsamer Lebenszyklus-Kontrollrahmen für moderne Pricing- und KI-Anwendungen. Dieser umfasst insbesondere Zweck und Scope der Anwendung, Rollen und Freigaben, Daten- und Modellversionierung, fachliche und technische Tests, Protokollierung produktiver Läufe, Monitoring, Incident-Prozesse, geregelte Änderungen sowie kontrollierte Stilllegung oder Ablösung. Viele Nachweise wie etwa Versionierung, Testevidenz, Freigaben, Logs und Monitoring-Ergebnisse können dabei sowohl für DORA als auch für den AI Act genutzt werden.

Zusätzliche Anforderungen ergeben sich je nach Regime: Der AI Act verlangt bei Hochrisiko-KI insbesondere Konzepte zur menschlichen Aufsicht, Daten-Governance, Transparenz und technischen Dokumentation. DORA ergänzt dies um Anforderungen an Betriebsresilienz, Wiederanlauf, IKT-Incident-Management sowie das Drittparteien- und Cloud-Risikomanagement. Inhaltliche Widersprüche zwischen beiden Regimen sind selten; mögliche Spannungsfelder, etwa zwischen Transparenz und Sicherheitsprinzipien oder zwischen Logging und Datenminimierung, sollten durch rollenbasierte Zugriffe, abgestufte Transparenz sowie klare Aufbewahrungs- und Löschkonzepte gelöst werden.

3.4. Anforderungen an Daten und Modelle in der Krankenversicherung

3.4.1. Anforderungen im Rahmen der Datennutzung

Die Nutzung von KI und modernen Methoden erfolgt in aller Regel im Zusammenspiel mit einer großen Menge an Daten. In der Krankenversicherung stellen Daten allerdings einen sehr sensiblen Faktor dar; bei ihrer Nutzung ist daher große Sorgfalt geboten. Die aktuariellen Grundsätze zum Umgang mit (Gesundheits-)Daten – von Datenherkunft und Compliance über personenbezogene und besonders schützenswerte Daten bis zur Diskriminierungsvermeidung – vertieft der DAV-Ergebnisbericht „Umgang mit Daten im Bereich Data Science“ des Ausschusses ADS / AI vom 14.02.2020.[9]

In der aktuariellen Arbeit in der Krankenversicherung sind folgende Prinzipien zu beachten:

- **Allgemeine Prinzipien im Umgang mit Daten und Compliance:** Die Herkunft der Daten und deren Verwendung sind innerhalb der eigenen Prozesse transparent und nachvollziehbar zu dokumentieren. Annahmen, Korrelationen und mögliche Ursachen sowie Unsicherheiten und Einschränkungen in den Ergebnissen müssen klar kommuniziert werden.

[9] Siehe Ergebnisbericht „Umgang mit Daten im Bereich Data Science“ des Ausschusses ADS / AI in der Fassung vom 14.02.2020.

- **Umgang mit personenbezogenen Daten:** Nach der Datenschutz-Grundverordnung (DSGVO) dürfen Daten nur für bekannte, rechtlich zulässige Zwecke verwendet werden, wobei eine sorgfältige Dokumentation erforderlich ist. Personenbezogene Daten sollen nur in dem Maße gespeichert und verarbeitet werden, wie es für den Zweck notwendig ist. Wo möglich und sinnvoll, sollten Daten anonymisiert oder pseudonymisiert werden.
- **Besondere Kategorien personenbezogener Daten (z. B. Gesundheitsdaten):** Die Erhebung und Verarbeitung sensibler personenbezogener Daten, wie Gesundheitsdaten, darf nur erfolgen, wenn eine gesetzliche Grundlage besteht oder eine ausdrückliche Einwilligung vorliegt. Ohne ausdrückliche Zustimmung sind Gesundheitsdaten nur dann verwendbar, wenn sie bereits anonymisiert sind.
- **Vermeidung von Diskriminierung:** Diskriminierung, etwa aufgrund von geschützten Merkmalen, muss in Kalkulations- und Bewertungsverfahren vermieden werden – sowohl unmittelbare (direkte) als auch mittelbare (indirekte) Diskriminierung. Differenzierende Kriterien dürfen nur verwendet werden, wenn sie nach dem Allgemeinen Gleichbehandlungsgesetz (AGG) zulässig sind und auf anerkannten risikoadäquaten Prinzipien beruhen. Aktuarinnen und Aktuare müssen prüfen, welche Variablen in Modellen verwendet werden und sicherstellen, dass keine unzulässige Diskriminierung entsteht – sowohl präventiv bei der Modellierung als auch ex post durch Plausibilisierung und Validierung.

Die Anwendung von KI und immer komplexeren Methoden birgt ein hohes Potenzial für Verbesserungen, zugleich aber auch das Risiko, dass vorliegende Verzerrungen unbeabsichtigt verstärkt werden und zu ungewünschten Diskriminierungen führen. Um dies zu vermeiden oder mindestens zu reduzieren, sind Maßnahmen zu ergreifen. Diese können technischer oder nicht-technischer Art sein, um einen verantwortungsvollen Umgang mit Bias und Diskriminierung sicherzustellen. Näheres dazu enthält der DAV-Ergebnisbericht „Vermeidung von Bias und Diskriminierung im Bereich Data Science“.[10]

3.4.2. Anforderungen an Modelle und Erklärbarkeit

Eine wichtige Anforderung an Modelle stellt die Erklärbarkeit dar. Diese gestaltet sich zunehmend komplizierter im Zusammenhang immer komplexerer Anwendungen einschließlich der Anwendung von KI-Tools. Erklärbarkeit ist ein vielschichtiges Konzept, für das keine eindeutige Begriffsdefinition existiert. Es können prinzipienbasierte Kriterien definiert werden, wann ein Modell als ausreichend erklärbar gilt. Der erforderliche Grad der Erklärbarkeit hängt dabei von der Anwendung und den Adressaten ab. Bei der Anwendung von Erklärbarkeitsmethoden ist aktuarielle Sorgfalt anzuwenden. Näheres dazu enthält der DAV-Ergebnisbericht „Explainable Artificial Intelligence: Ein aktueller Überblick für Aktuarinnen und Aktuare“.[11]

Neben der Erklärbarkeit ist die Fairness der Modelle eine zentrale Anforderung. Modelle sollen vergleichbare Sachverhalte vergleichbar behandeln und dürfen keine unzulässige unmittelbare oder mittelbare Diskriminierung erzeugen; hierzu sind geeignete Fairness-Kriterien zu definieren und sowohl bei der Entwicklung als auch im laufenden Betrieb zu überprüfen. Die methodischen Grundlagen der Erklärbarkeit sind in Anhang A (Abschnitt A.3.2) dargestellt.

[10] Siehe Ergebnisbericht „Vermeidung von Bias und Diskriminierung im Bereich Data Science“ des Ausschusses ADS / AI in der Fassung vom 11.03.2025.

[11] Siehe Ergebnisbericht „Explainable Artificial Intelligence: Ein aktueller Überblick für Aktuarinnen und Aktuare“ des Ausschusses ADS / AI in der Fassung vom 27.05.2024.

4. Einsatz von künstlicher Intelligenz und modernen Methoden in Kernprozessen der privaten Krankenversicherung

Dieses Kapitel zeigt anhand konkreter Anwendungsfälle, an welchen Stellen der aktuariellen Kernprozesse – insbesondere im Kalkulationsablauf der Beitragsanpassung und bei der Analyse unternehmenseigener Daten – künstliche Intelligenz und moderne Methoden sinnvoll eingesetzt werden können.

4.1. Anwendungsfälle im Kalkulationsablauf der Beitragsanpassung

Die Beitragsanpassung in der privaten Krankenversicherung ist ein vielschichtiger Prozess, der von der Datenaufbereitung über die Kalkulation bis zur Limitierung und Dokumentation reicht und in dem Berechnungs-, Bewertungs- und Entscheidungsschritte aufeinander aufbauen. Entlang dieses Ablaufs ergeben sich an mehreren Stellen Ansatzpunkte für den sinnvollen Einsatz von KI und modernen Methoden. In den folgenden Abschnitten werden Möglichkeiten der Anwendung innerhalb des Kalkulationsablaufs systematisch aufgezeigt und jeweils kurz erläutert.

Betrachtet werden die drei in Kapitel 2 eingeführten Kategorien – KI-Anwendungen, moderne Methoden und Automatisierungstools (siehe die Abschnitte 2.1 bis 2.3):

- **KI-Anwendungen:** Ziel ist die Beschleunigung oder Qualitätsverbesserung durch Mustererkennung, generative Sprachverarbeitung oder die Verarbeitung großer, heterogener Datenmengen. Dies kann durch den Einsatz von KI-Anwendungen, d. h. insbesondere Verfahren des maschinellen Lernens und LLMs, erreicht werden. Auch der Einsatz von KI-Agenten kann hierunter fallen.
- **Moderne Methoden:** Ergänzung oder Ersatz klassischer Verfahren durch erweiterte aktuariell-statistische Modellbildung, datengetriebene Stützungsverfahren, Simulationsmodelle oder aussagekräftige Dashboards. Es werden insbesondere Qualitätsverbesserungen und verbesserte Entscheidungsgrundlagen generiert.
- **Automatisierungstools:** Ziel ist hierbei, Verbesserungen der Prozessdurchlaufzeiten und -zuverlässigkeit durch klassische, deterministische Automatisierung ohne eigene KI-Komponente zu erreichen. Hierunter fällt insbesondere die Ablösung manueller (Excel-)Arbeit durch in zeitgemäßen Programmiersprachen entwickelte Softwarelösungen.

Die Kategorien überschneiden sich begrifflich – Automatisierungstools sind nach Abschnitt 2.3 eine Teilmenge der modernen Methoden. Für die folgende Übersicht wird jeder Ansatz derjenigen Kategorie zugeordnet, die ihn am treffendsten beschreibt.

In Tabelle 1 sind typische Schritte des Kalkulationsablaufs im Rahmen einer Beitragsanpassung zusammengefasst. Die Darstellung gliedert den Kalkulationsablauf in vier Phasen: Vorbereitung, Kalkulation der Grundkopfschäden, Kalkulation der Rechnungsgrundlagen sowie Limitierung und Dokumentation, ergänzt um das Querschnittsthema der Ablösung von Altsystemen durch eine moderne Kalkulationsumgebung.

Innerhalb der vier Phasen werden die einzelnen Schritte beschrieben und mögliche Ansatzpunkte aufgezeigt, die jeweils einer der drei genannten Kategorien zugeordnet sind. Die Aufstellung ist beispielhaft und erhebt keinen Anspruch auf Vollständigkeit.

Grundsätzlich lassen sich viele der genannten KI-Anwendungen auch als deterministisch arbeitendes Programm umsetzen. Auch dann kann der Einsatz von KI sinnvoll sein, weil sie – insbesondere als LLM – den Programmcode deutlich schneller erstellt als eine rein manuelle Entwicklung, während das produktive Programm selbst deterministisch bleibt (vgl. Abschnitt 3.3). Für alle drei Kategorien gelten dieselben aufsichtsrechtlichen Leitplanken, wie sie in Kapitel 3 dargestellt sind (insbesondere die Abschnitte 3.1, 3.2 und 3.4). Die beschriebenen Verfahren sind optionale Möglichkeiten und keine verpflichtenden Standards; der aktuarielle Gestaltungsspielraum bleibt unberührt.

Zwei der genannten Anwendungen werden anschließend vertieft: Abschnitt 5.1 stellt eine bereits umgesetzte neue Architektur des Limitierungsmodells vor (Schritt 10), Abschnitt 5.2 eine prototypisch erprobte LLM-gestützte Validierung Technischer Berechnungsgrundlagen (Schritt 11). In der folgenden Übersicht werden beide Themen nur kurz eingeordnet.

Schritt	Kategorie	Ansatz zur Optimierung
Phase I: Vorbereitung		
1 Ermittlung der Auslösenden Faktoren	KI-Anwendungen	Anomalie-Erkennung in Schadendaten
	KI-Anwendungen	Gestützte Kommentierung mit Hilfe von Natural Language Processing (NLP)
2 Datenselektion	KI-Anwendungen	Simulation von Stützdaten bzw. fiktiver Stufen
	Moderne Methoden	Feature Engineering von Risikomerkmalen
	Automatisierungstools	Automatische Datenqualitätsprüfung
Phase II: Kalkulation – Grundkopfschaden		
3 Datengrundlage: Trendermittlung und medizinische Inflation	KI-Anwendungen	Prognose der Bestandsentwicklung nach Diagnosen
	Moderne Methoden	Zeitreihenmodelle ((S)ARIMA, State-Space-Modelle)
	Moderne Methoden	Statistische Erkennung von Ausreißerjahren
	Automatisierungstools	Integration externer Datenquellen
4 Profilfestlegung	Moderne Methoden	Generalisierte Additive Modelle (GAMs)
	Moderne Methoden	ML-Verfahren als Plausibilisierungsbenchmark
	Moderne Methoden	Stützungsverfahren mit datengetriebener Kredibilität
	Moderne Methoden	Kreuzvalidierung zur Modellselektion
	Automatisierungstools	Automatische Erkennung nicht mehr aussagekräftiger Daten
5 Festlegung der Grundkopfschäden	KI-Anwendungen	Begründungstexte mit Hilfe von LLMs
	Moderne Methoden	Stützungsverfahren mit Shrinkage
Phase III: Kalkulation – Rechnungsgrundlagen		
6 Sterbetafel	–	Aktuell keine Optimierungsoption ersichtlich
7 Stornotafel	Moderne Methoden	Bestandsabhängige Stornomodellierung
8 Kosten	–	Aktuell keine Optimierungsoption ersichtlich
9 Beitragsberechnung	–	Aktuell keine Optimierungsoption ersichtlich

Schritt	Kategorie	Ansatz zur Optimierung
Phase IV: Limitierung und Dokumentation		
10 Limitierung	Moderne Methoden	Optimierung des RfB-Mitteleinsatzes
	Moderne Methoden	Simulationsmodelle anstelle statischer Limitierungsregeln
	Automatisierungstools	Neue Architektur des Limitierungsmodells
11 Erstellung und Validierung Technischer Berechnungsgrundlagen und Dokumentationen	KI-Anwendungen	Erstellung von TBs
	KI-Anwendungen	Validierung von TBs
12 Kommunikation	KI-Anwendungen	Zusammenfassungen für unterschiedliche Adressaten
	Automatisierungstools	Automatisierte Aufbereitung der Treuhänder-Unterlagen
	Automatisierungstools	Unterstützung bei Beitragsanpassungsklagen
Querschnittsthema		
Ablösung von Altsystemen durch eine moderne Kalkulationsumgebung	Automatisierungstools	Softwaregestütztes Kalkulationstool
	Moderne Methoden	Passgenauere Frontends
	Moderne Methoden	Versionierung und Reproduzierbarkeit
	Moderne Methoden	Automatisierte Tests

Tabelle 1: Übersicht von Verbesserungspotenzialen entlang des Kalkulationsablaufs einer Beitragsanpassung.

Im Folgenden werden die zwölf nummerierten Schritte des Kalkulationsablaufs sowie das Querschnittsthema einzeln betrachtet. Zu jedem Schritt werden der Zweck kurz beschrieben und die Ansatzpunkte je Kategorie skizziert.

4.1.1. Phase I: Vorbereitung

Schritt 1 – Ermittlung der Auslösenden Faktoren (AF-Ermittlung)

Zweck

Bestimmung der Auslösenden Faktoren Schaden und Sterblichkeit als rechtliche Voraussetzung einer Beitragsanpassung gemäß § 203 Absatz 2 VVG in Verbindung mit § 155 VAG sowie §§ 15 und 16 KVAV. Vor der AF-Ermittlung sind die Schadendaten auf Qualität und Sondervorgänge hin zu prüfen.

Anwendungen zur Optimierung

KI-Anwendungen

- **Anomalie-Erkennung in Schadendaten:** Verfahren des unüberwachten Lernens – etwa Isolation Forest oder Autoencoder – markieren auffällige Einzelschäden, Bestände oder Strukturbrüche vor der AF-Ermittlung. Gegenüber der heute häufig regelbasierten Vorgehensweise mit fixen absoluten, relativen oder quantilsbasierten Schwellenwerten für Groß-

und Extremschäden ergeben sich eine gleichmäßigere Erkennung außergewöhnlicher Schäden und systematischer Sondervorgänge sowie eine bessere Vergleichbarkeit. Eine ähnliche Erkennung ist auch für die Datenselektion (Schritt 2) anwendbar.

- **Gestützte Kommentierung mit Hilfe von Natural Language Processing (NLP) zur Abstimmung mit der mathematischen Treuhänderin oder dem mathematischen Treuhänder:** Automatische Kategorisierung von Freitext-Bemerkungsfeldern (Schadensart, Sondersachverhalt, Schwangerschafts- und Mutterschaftsfälle) etwa durch ein LLM. Ziel ist es, bei der Abstimmung mit der Treuhänderin oder dem Treuhänder Sondersachverhalte schneller auffindbar und nachvollziehbar begründbar zu machen. Die Anwendung ist dort sinnvoll möglich, wo Freitextfelder über lange Zeiträume in vergleichbarer Qualität gepflegt vorliegen.

Schritt 2 – Datenselektion

Zweck

Aufbereitung der Bestands- und Schadendaten für die folgenden Kalkulationsschritte: Bestände bestimmen, Beginnjahre zur Eliminierung von Wartezeit- und Selektionseffekten festlegen, Korrekturen aus Leistungsänderungen entsprechend den Festlegungen in den Technischen Berechnungsgrundlagen anwenden (§ 15 Absatz 2 KVAV), risikoadäquate Merkmale ableiten.

Anwendungen zur Optimierung

KI-Anwendungen

- **Simulation von Stützdaten bzw. fiktiver Stufen mit Hilfe von maschinellem Lernen:** Erzeugung von Bestandsverteilungen für Beobachtungseinheiten, die in der Realität zu klein für eine eigenständige Auswertung sind, beispielsweise bei der Einführung neuer Tarifvarianten. Methodisch ist diese Anwendung als Stützungsverfahren einzuordnen; sie unterliegt deshalb den Begründungs- und Validierungsanforderungen, die für jedes Stützverfahren gelten. (Vgl. § 15 Absatz 4 und § 6 Absatz 3 KVAV sowie den DAV-Hinweis zu „Kalkulation und Bestandsgröße in der privaten Krankenversicherung“ [12].)

Moderne Methoden

- **Feature Engineering von Risikomerkmalen:** Konstruktion erklärungsstarker Merkmale, z. B. Komorbiditätsindikatoren, regionale Cluster, dynamische Risikoindikatoren, die die nachfolgende Profil- und Stützungsmodellierung verbessern. Nutzung im Rahmen von Machine-Learning-Modellen. Bei der Auswahl der Merkmale ist auf die Diskriminierungsfreiheit zu achten, insbesondere mit Blick auf das Unisex-Gebot und das AGG. Die Merkmale dienen als Eingangsgrößen der Profil- und Stützungsmodellierung sowie als Risikoindikatoren für die Plausibilisierung.

Automatisierungstools

- **Automatische Datenqualitätsprüfung:** Reproduzierbar laufende Prüfroutine (Wertebereichs- und Konsistenzprüfungen, Dublettenerkennung, regelbasierte Plausibilisierung) vor jeder Kalkulation. Implementiert als automatisierter Schritt (z. B. innerhalb einer versionierten Pipeline-Komponente, d. h. eines automatisierten, unter Versionskontrolle stehenden Verarbeitungsschritts), deren Resultate auditierbar protokolliert werden. Diese Maßnahme unterstützt die geordnete Nachvollziehbarkeit nach § 9 KVAV und bildet eine sinnvolle Grundlage für weitergehende KI-Anwendungen. Zugleich unterstützt sie die DORA-Anforderungen an Änderungs-, Test- und Nachweisprozesse (vgl. Abschnitt 3.2).

[12] Siehe Hinweis „Kalkulation und Bestandsgröße in der privaten Krankenversicherung“ des Ausschusses Krankenversicherung in der Fassung vom 21.06.2023.

4.1.2. Phase II: Kalkulation – Grundkopfschaden

Schritt 3 – Datengrundlage: Trendermittlung und medizinische Inflation

Zweck

Ermittlung von Trends in den verwendeten Bestands- und Leistungsdaten, mit denen die Leistungen auf den Kalkulationszeitpunkt fortgeschrieben werden – gegebenenfalls nach Tarif(gruppe), Leistungsbereich, etc. getrennt.

Anwendungen zur Optimierung

KI-Anwendungen

- **Prognose der Bestandsentwicklung nach Diagnosen mit Hilfe von Machine-Learning-Modellen:** Vorhersage zukünftiger Schadenniveaus über die Modellierung der Diagnose-Mix-Entwicklung im Bestand anstelle aggregierter Schadenzeitreihen. Sie wird hier als perspektivisches Forschungsfeld benannt und nicht in Gänze ausgeführt; zu weiteren Prognoseanwendungen siehe Abschnitt 4.2, zur vertieften Nutzung von Diagnosedaten in der Schadenbedarfsprognose Abschnitt 5.3. Voraussetzung ist eine verlässliche Diagnoseschlüsselqualität in den Schadendaten.[13]

Moderne Methoden

- **Zeitreihenmodelle:** Modellierung mit expliziter Trend-, Saison- und Sondereffekt-Komponente in den Bestands- und Leistungsdaten mittels struktureller Zeitreihen-/State-Space-Modelle (z. B. (S)ARIMA) – etwa Isolation der COVID-19-Anomalien der Jahre 2020 und 2021. Strukturierte Modelle erleichtern die Begründung der Trendannahmen gegenüber der Treuhänderin oder dem Treuhänder im Vergleich zu rein linearen Regressionen. Die so ermittelten Trends gehen in die Festlegung der Grundkopfschäden ein (Schritt 5).
- **Statistische Erkennung von Ausreißerjahren:** Identifikation auffälliger Beobachtungsjahre mit statistischen Verfahren, etwa mittels Verfahren des unüberwachten maschinellen Lernens, und daraus abgeleitete Vorschläge für die Trendermittlung, die anschließend aktuariell zu bewerten sind.

Automatisierungstools

- **Integration externer Datenquellen:** Anbindung beispielsweise an Statistiken des Statistischen Bundesamtes, des PKV-Verbandes, der Kassenzahnärztlichen Bundesvereinigung sowie weiterer Branchenkennzahlen als Plausibilisierungsbenchmark. Die Anbindung soll als versionierte Datenpipeline ausgeführt sein, damit der Datenstand jeder Kalkulation reproduzierbar bleibt.

Schritt 4 – Profilfestlegung

Zweck

Ermittlung des Alters- und Geschlechtsprofils, das den Kopfschadenverlauf strukturiert. Profile sind über Tarife in der Regel stabil und werden nicht jährlich angepasst.

Anwendungen zur Optimierung

Moderne Methoden

- **Generalisierte Additive Modelle (GAMs)** (ggf. mit Monotonie- bzw. Konvexitäts-Constraints): Datengetriebene, glatte Profile mit fachlich erwünschten bzw. geforderten Eigenschaften. Die Constraints sichern die erwünschten bzw. geforderten Eigenschaften (z. B. die Mo-

[13] Hinweis: Datenschutzrechtlich sind Diagnosedaten der besonderen Kategorie nach Artikel 9 DSGVO zuzuordnen.

notonie des Beitragsverlaufs) und verhindern unplausible Ausschläge in dünn besetzten Altersbereichen. GAMs eignen sich für den produktiven Einsatz, da sie interpretierbar, formkontrollierbar und reproduzierbar sind.

- **ML-Verfahren als Plausibilisierungsbenchmark:** Gradient-Boosting-Modelle oder Random Forests dienen als Vergleichsmaßstab zum klassisch ermittelten Profil. Differenzen zwischen Benchmark und Profil werden aktuariell analysiert und interpretiert. Gradient Boosting und Random Forests dienen wegen ihres Black-Box-Charakters nur als Vergleichsmaßstab.
- **Stützungsverfahren mit datengetriebener Kredibilität:** Statt der heute oft konventionsbasierten Gewichtung wird die Stützungsgewichtung über Out-of-Sample-Validierung optimiert. Methodisch verwandt mit dem Bühlmann-Straub-Schätzer in einer modernen, datengetriebenen Ausprägung. Die Gewichtung wird damit nicht fest vorgegeben, sondern datengetrieben aus der Validierungsgüte abgeleitet und ist gegenüber der Treuhänderin oder dem Treuhänder reproduzierbar begründbar.
- **Kreuzvalidierung zur Modellselektion:** Verlagerung der Wahl von Glättungs- und Stützungsparametern vom Erfahrungsurteil zu einer datengestützten Auswahl mit reproduzierbarer Begründung gegenüber der Treuhänderin oder dem Treuhänder.

Automatisierungstools

- **Automatische Erkennung nicht mehr aussagekräftiger Daten:** Prüfung auf zu kleine Bestände (vgl. den DAV-Hinweis zu „Kalkulation und Bestandsgröße in der privaten Krankenversicherung“[14]), unzureichend besetzte Alterskohorten sowie eine im Jahresvergleich zunehmende Varianz.

Schritt 5 – Festlegung der Grundkopfschäden

Zweck

Festlegung des Niveaus der Grundkopfschäden pro Beobachtungseinheit, nach Anwendung von Trend und Stützung.

Anwendungen zur Optimierung

KI-Anwendungen

- **Begründungstexte mit Hilfe von LLMs:** Automatische Erstellung der Begründungstexte zur Grundkopfschaden-Festlegung als Entwurfshilfe. Methodisch eng verwandt mit der LLM-gestützten TB-Erstellung (siehe Abschnitt 5.2).

Moderne Methoden

- **Stützungsverfahren mit Shrinkage:** Gewichtete Kombination von Eigenbestand und Stützniveau (z. B. Stütztarif, Profil × Tarifschaden) mit kredibilitätsabhängiger Gewichtung. Bei mehreren Stützen ist eine optimale Gewichtung über Validierung möglich. Die Methodik führt zu einer verbesserten Prognosequalität.

[14] Siehe Hinweis „Kalkulation und Bestandsgröße in der privaten Krankenversicherung“ des Ausschusses Krankenversicherung in der Fassung vom 21.06.2023.

4.1.3. Phase III: Kalkulation – Rechnungsgrundlagen

Schritt 6 – Sterbetafel

Zweck

Maßgebend bleibt die jeweils aktuelle PKV-Sterbetafel; nach dem DAV-Hinweis zu Technischen Berechnungsgrundlagen ist stets die aktuellste Tafel zugrunde zu legen und ihr Ansatz zu begründen.[15] Eigene Sterbetafeln sind nur in begründeten Sonderfällen einsetzbar; ihre Begründung gegenüber der Treuhänderin oder dem Treuhänder ist anspruchsvoll.

Anwendungen zur Optimierung

Aktuell keine Optimierungsoption mit Hilfe von modernen Methoden oder KI ersichtlich.

Schritt 7 – Stornotafel

Zweck

Analyse von Treibern des Stornoverhaltens zur optimierten Festlegung der kalkulatorischen Stornoannahmen.

Anwendungen zur Optimierung

Moderne Methoden

- **Bestandsabhängige Stornomodellierung:** Modellierung des Stornos als Funktion mehrerer Treiber – Eintrittsalter, Versicherungsdauer, Tariftyp, Beitragshöhe und Beitragsanpassungshistorie, statt rein altersabhängiger Modellierung. Die Wirkung auf die Beitragsanpassung und die Limitierung im Bestand ist substanziell, weil die Stornoannahme den Limitierungseffekt mitbestimmt. Bei Tarifen der Neuen Welt ist zusätzlich die Reduktion der Stornowahrscheinlichkeiten zur Berücksichtigung des Übertragungswertes auszuweisen.

Schritt 8 – Kosten

Zweck

Die klassische Kostenrechnung gemäß § 8 Absatz 2 bis 4 KVAV bleibt üblich. Eine Activity-based-Cost-Allokation der Verwaltungs-, Schadenregulierungs- und Akquisitionskosten kann die Transparenz der Kostenherleitung erhöhen, insbesondere bei größeren Tarifportfolios.

Anwendungen zur Optimierung

Aktuell keine Optimierungsoption mit Hilfe von modernen Methoden oder KI ersichtlich.

Schritt 9 – Beitragsberechnung

Zweck

Die Beitragsberechnung ist ein deterministischer, etablierter und regulatorisch stark festgelegter Prozess. Verbesserungspotenziale liegen vorrangig bei der Datenbeschaffung (siehe oben) und der Implementierungsqualität der Kalkulationsmodelle und damit im Querschnittsthema zur Ablösung von Altsystemen (siehe Abschnitt 4.1.5).

Anwendungen zur Optimierung

Aktuell keine Optimierungsoption mit Hilfe von modernen Methoden oder KI ersichtlich.

[15] Siehe Hinweis „Erstellung und Inhalte Technischer Berechnungsgrundlagen in der privaten Krankenversicherung“ des Ausschusses Krankenversicherung in der Fassung vom 27.01.2021, Abschnitt 2.2.1.

4.1.4. Phase IV: Limitierung und Dokumentation

Schritt 10 – Limitierung

Zweck

Begrenzung der Beitragssprünge im Bestand unter Einsatz von RfB-Mitteln nach § 150 VAG. Die Festlegung der Limitierung erfolgt mit Zustimmung der Treuhänderin oder des Treuhänders.

Anwendungen zur Optimierung

Moderne Methoden

- **Optimierung des RfB-Mitteleinsatzes:** Anstelle einer heuristischen Mittelallokation kommt ein zielgrößengeführtes Optimierungsmodell zum Einsatz, das die Mittelverteilung unter expliziten Nebenbedingungen (Mittelbilanz, Mindestbeitragshöhe, Gleichbehandlungsgrundsatz innerhalb von Zielgruppen) löst. Das Modell kann mittels linearer Programmierung, Mixed Integer Programming oder stochastischer Programmierung erstellt werden.
- **Simulationsmodelle anstelle statischer Limitierungsregeln:** Auch klassische Monte-Carlo-Simulationen der Bestandsentwicklung können zur Entwicklung alternativer Limitierungsszenarien und zur Visualisierung der Verteilung zukünftiger Beitragspfade herangezogen werden. Dies erleichtert die Treuhänder-Diskussion über die Wirkung verschiedener Limitierungsregeln wesentlich.

Automatisierungstools

- **Neue Architektur des Limitierungsmodells:** Die bisher häufig in Altsystemen oder Excel implementierte Limitierungslogik kann durch eine modulare Architektur ersetzt werden. Dabei ist diese in einer modernen Programmiersprache umgesetzt und produktiv vertestbar: Es lässt sich damit eine deutliche Performance-Verbesserung erzielen. Eine eigene Bedienoberfläche (GUI) erlaubt der Aktuarin bzw. dem Aktuar die parametrisierte Steuerung der Läufe. Eine ausführliche Beschreibung einer möglichen modernen Architektur samt Wirkungsanalyse erfolgt in Abschnitt 5.1.

Schritt 11 – Erstellung und Validierung Technischer Berechnungsgrundlagen und Dokumentationen

Zweck

Erstellung und Fortschreibung der Technischen Berechnungsgrundlagen (TBs) sowie weiterer Dokumentation für die Treuhänder-Kommunikation. Der Inhalt von TBs ist Grundlage der Treuhänder-Prüfung und der Bafin-Aufsicht und die Struktur von TBs orientiert sich an den Vorgaben des entsprechenden DAV-Hinweises[16].

Anwendungen zur Optimierung

KI-Anwendungen

- **Erstellung von TBs:** Bewährte klassische Automatisierung bleibt dort erhalten, wo sie bereits trägt, etwa beim Auslesen der Zahlentabellen aus der Kalkulationsdatenbank. LLMs entfalten ihren Nutzen vor allem bei den narrativen Teilen, etwa bei Methoden- und Begründungstexten, sowie beim Entwurf einzelner TB-Abschnitte aus Vorgängerversionen und den strukturierten Eingabewerten der aktuellen Beitragsanpassung. Perspektivisch können Agentensysteme den vollständigen Entwurf einer TB übernehmen.

[16] Siehe Hinweis „Erstellung und Inhalte Technischer Berechnungsgrundlagen in der privaten Krankenversicherung“ des Ausschusses Krankenversicherung in der Fassung vom 27.01.2021.

- **Validierung von TBs:** Zweistufige Prüfung einer entworfenen oder bestehenden TB aus regelbasierten Konsistenz- und Vollständigkeitsprüfungen – etwa zu Querverweisen, Symbolverzeichnis, Änderungshistorie sowie Rahmen-TB- und AVB-Verweisen – und ergänzenden LLM-gestützten Prüfschritten für sprachlich und inhaltlich zu beurteilende Zusammenhänge. Der Ansatz ist in Abschnitt 5.2 ausführlich beschrieben.

Schritt 12 – Kommunikation

Zweck

Aufbereitung der Kalkulationsergebnisse für unterschiedliche Zielgruppen, etwa die Verantwortlichen Aktuarinnen und Aktuare, die Treuhänderinnen und Treuhänder und die Versicherten.

Anwendungen zur Optimierung

KI-Anwendungen

- **Zusammenfassungen für unterschiedliche Adressaten:** Mit Hilfe von LLMs werden die Kalkulationsergebnisse adressatengerecht aufbereitet – technisches Detail (z. B. detaillierte Ergebnisse, Parameteränderungen, methodische Hinweise) für die Verantwortlichen Aktuarinnen und Aktuare, eine regulatorisch-vergleichende Darstellung für die Treuhänderinnen und Treuhänder sowie laienverständliche Begründungsbausteine für die Versicherten-Information und für die Vermittler, die entsprechende Fragen im Kundengespräch beantworten. Agentensysteme können die dafür benötigten Kennzahlen zusätzlich selbstständig aus der Kalkulationsstrecke zusammentragen.

Automatisierungstools

- **Automatisierte Aufbereitung der Treuhänder-Unterlagen:** Das Treuhänder-Paket – bestehend aus Technischen Berechnungsgrundlagen, Anlagen, Beispielrechnungen und Limitierungsbericht – wird automatisch softwaregetrieben aus der Kalkulationsstrecke generiert, statt manuell zusammengetragen. Konsistenz zwischen den Dokumenten ist damit konstruktiv sichergestellt. Die Prüfung und Freigabe des Pakets verbleibt beim Fachbereich (Human-in-the-Loop).
- **Unterstützung bei Beitragsanpassungsklagen:** Automatisierter Abgleich und Zusammenstellung der für ein Verfahren benötigten Angaben – Beitragszahlungen sowie Anpassungen und deren Begründungen der vergangenen Jahre – zu einem einheitlichen Informationspaket für Gericht und beteiligte Anwaltskanzleien. Da die Anfragen weder standardisiert eingehen noch einheitliche Angaben abfragen, war diese Zusammenstellung bislang aufwendige Handarbeit. Die Prüfung und Freigabe des Pakets verbleibt beim Fachbereich (Human-in-the-Loop).

4.1.5. Querschnittsthema: Ablösung von Altsystemen durch eine moderne Kalkulationsumgebung

Zweck

Das Querschnittsthema bildet die infrastrukturelle Grundlage der vorgenannten Anwendungen: Erst eine reproduzierbare, versionierte und getestete Kalkulationsumgebung erfüllt die Nachvollziehbarkeitspflicht nach § 9 KVAV und die Anforderungen des AI Acts und der DORA an Dokumentation, Logging und Reproduzierbarkeit.

Anwendungen zur Optimierung

Automatisierungstools

- **Softwaregestütztes Kalkulationstool:** Implementierung in Python, R oder einer anderen modernen Programmiersprache mit klarer Trennung der Module Daten, Rechnungsgrundlagen, Beitragsberechnung und Limitierung. Vektorisierte Berechnung über alle Beobach-

tungseinheiten erlaubt eine Beschleunigung etwa gegenüber zellbasierten Excel-Implementierungen. Für die Umsetzung eignen sich moderne Datenplattformen mit nativer R-/Python-Integration, Versionierung, automatisierten Tests und Governance-Unterstützung.

Moderne Methoden

- **Passgenauere Frontends:** Web- oder Desktop-Bedienoberflächen, die parametrisierte Kalkulationsläufe und Treuhänder-Demonstrationen unterstützen und die Ergebnisse klarer visualisieren, als es in gewachsenen Excel-Tabellen möglich ist. Das reduziert zugleich die Fehleranfälligkeit.
- **Versionierung und Reproduzierbarkeit:** Versionskontrolle des Codes (z. B. mit Git oder Subversion), versionierte Datenartefakte (etwa über Object Storage), Model-Cards für ML-Komponenten; beim Einsatz von LLMs sind zusätzlich der Versionsbezeichner des Modells und die verwendeten Hyperparameter wie die Temperature festzuhalten. Jede produktive Kalkulation muss exakt reproduzierbar sein. Dies ist insbesondere bei einer Prüfung durch die Aufsicht oder einer Beitragsanpassungsklage von Bedeutung.
- **Automatisierte Tests:** Unit-Tests auf mathematische Identitäten (z. B. Kommutationswert-Beziehungen, Äquivalenzprinzip, Prämienberechnung) und Property-Tests auf Eigenschaften wie die Monotonie der Profile. Ergänzend werden bei jeder Änderung die Ergebnisse eines Referenztarifs vollständig neu berechnet und mit einem festgehaltenen Sollstand verglichen (Regressionstest). So fallen unbeabsichtigte Nebenwirkungen einer Korrektur sofort auf.

4.2. Anwendungsfälle von modernen Methoden und KI zur Analyse unternehmenseigener Daten und zur Ableitung von Bestandsprognosen

Das maschinelle Lernen hat sich in den letzten Jahren als zentraler Ansatz für datengetriebene Analysen und Prognosen etabliert. Ziel ist es, aus historischen Daten Vorhersagemodelle abzuleiten, die auch auf neue, unbekannte Beobachtungen verlässlich angewendet werden können. Insbesondere im aktuariellen Kontext eröffnen sich dadurch neue Möglichkeiten, komplexe Zusammenhänge in großen und heterogenen Datensätzen zu modellieren. Solche Analysen und Modelle lassen sich als Business-Intelligence-Komponente in die übergreifende Datenlandschaft einbetten.

Im Fokus dieses Abschnitts stehen ausgewählte Prognoseverfahren, die sich für tabellarische Daten als besonders leistungsfähig erwiesen haben. Dazu zählen klassische Methoden wie die logistische Regression ebenso wie seit einigen Jahren etablierte Ansätze auf Basis von Entscheidungsbäumen und Gradient Boosting; Abschnitt A.2 stellt diese Verfahren kurz vor.

In der Vergangenheit erforderte die Implementierung dieser Ansätze neben tiefem aktuariellen Verständnis für die zu untersuchenden Sachverhalte auch profunde Programmier- und Modellierungskenntnisse; das stellte in der Praxis eine nicht zu unterschätzende Hürde für ihre Verbreitung dar.

Ein besonderer Schwerpunkt liegt daher auf dem Einsatz moderner Tools, insbesondere von AutoML-Tools, die große Teile des Modellierungsprozesses automatisieren und auch komplexe Modelle effizient erzeugen können. Beispiele dafür sind im Abschnitt A.3.1 genannt.

Trotz der enormen Vereinfachungen bei der Nutzung moderner Methoden bleibt zentral in der Verantwortung von Aktuarinnen und Aktuaren, die Qualität der zu analysierenden Daten sicherzustellen, und durch geeignete Aufbereitung der Originaldaten (durch Bereinigung von Ausreißern, Umgang mit lückenhaften Informationen etc.) so weit zu optimieren, dass die verwendeten Methoden ihre Leistungsfähigkeit auch ausspielen können.

Gleichzeitig gewinnt die Erklärbarkeit von Modellen zunehmend an Bedeutung, da viele leistungsstarke Verfahren sogenannte Black-Box-Modelle darstellen. Geeignete Methoden ermöglichen es, Modellentscheidungen transparent zu machen und deren Einsatz im fachlichen Kontext zu unterstützen; einen Überblick gibt Abschnitt A.3.2.

Die Konzepte werden in Anhang A genauer erläutert und in den Abschnitten 5.4 und 5.5 durch zwei praxisnahe Anwendungsfälle illustriert:

- Im ersten Anwendungsfall (Abschnitt 5.4) wird eine Ereignisprognose zur Vorhersage von Krankenhauswiedereinweisungen mithilfe moderner ML-Verfahren umgesetzt.
- Der zweite Anwendungsfall (Abschnitt 5.5) zeigt anhand eines Krankenversicherungsportfolios, wie Daten explorativ analysiert und klassische Modellansätze zur Beschreibung von Prämienstrukturen eingesetzt werden können.

Die im vorherigen Abschnitt beschriebenen Anwendungen für KI-Systeme und moderne Methoden im Kontext der Beitragsanpassung (Abschnitt 4.1) lassen sich grundsätzlich auch für interne Analysen und für Prognosezwecke nutzen.

4.3. Weitere Anwendungsmöglichkeiten und Ausblick

Als moderne Methode zur Anreicherung der Datenbasis kommt zunehmend die Erzeugung synthetischer Daten in Betracht: Wenn unternehmenseigene Daten aus Gründen des Datenschutzes nicht oder nur eingeschränkt zu Analyse- oder Modellierungszwecken verwendet werden dürfen, etwa weil sie dafür auf eine Drittanbieter-Plattform geladen werden müssten, lassen sich Datensätze künstlich erzeugen, die den Originaldaten strukturell so ähnlich sind, dass sie sich als Ersatz-Datenbasis für die beabsichtigten Analysen und Modellierungen eignen.

Die DAV-Arbeitsgruppe „Synthetisierung von Daten“ befasst sich mit dieser Fragestellung. Sie analysiert Chancen, Risiken und regulatorische Rahmenbedingungen und entwickelt praxistaugliche Tools für die Erzeugung synthetischer Daten in der Versicherungswirtschaft. Erste Ergebnisse wurden auf der Herbsttagung 2025 von DAV und DGVFM vorgestellt; der Vortrag ist auf actuview[17] abrufbar.

In jüngerer Zeit wurden auch Ansätze entwickelt und veröffentlicht, um mit Methoden von großen Sprachmodellen aus individuellen Patientenhistorien Prognosen abzuleiten:

BEHRT (Bidirectional Encoder representation for Health Records using Transformers) ist ein auf Transformern basierendes Deep-Learning-Modell, das speziell für die Analyse elektronischer Gesundheitsakten zur Vorhersage von Patientenergebnissen und Krankheitsrisiken entwickelt wurde. Kernidee: So wie ein LLM die kontextuelle Bedeutung von Wörtern in Sätzen lernt, lernt BEHRT aus Folgen medizinischer Ereignisse (Diagnosen, Verschreibungen, Laborergebnisse) in der Gesundheitsakte eines Patienten.[18]

Inzwischen existiert auch eine deutsche Adaption von BEHRT, die speziell auf elektronische Gesundheitsakten in Deutschland zugeschnitten ist (GERBEHRT, entwickelt am Zentralinstitut für die kassenärztliche Versorgung in Deutschland, 2025).[19]

Diese Modelle befinden sich noch im Forschungsstadium und werden im vorliegenden Ergebnisbericht nicht weiter vertieft, können aber in Zukunft den aktuariellen „Baukasten“ bereichern.

[17] Siehe https://actuview.com/videos/erzeugung-synthetischer-daten-use-cases-software-und-anleitung-4921.

[18] Li, Y., Rao, S., Ayala Solares, J. R. et al. (2020): *BEHRT: Transformer for Electronic Health Records*, Scientific Reports 10, 7155, abrufbar unter https://doi.org/10.1038/s41598-020-62922-y

[19] Seidel, A., Steiger, E., von Samson-Himmelstjerna, F. A., Kroll, L. E. (2025): *GERBEHRT: A BERT-based Model Tailored for German Electronic Health Records – Potential in Chronic Kidney Disease Prediction*, Preprint, medRxiv, abrufbar unter https://doi.org/10.1101/2025.10.24.25338721

5. Vertiefung ausgewählter Anwendungsfälle

Die folgenden Abschnitte vertiefen fünf Anwendungsfälle aus der Arbeit der Arbeitsgruppe: ein Limitierungstool und eine Prüfstrecke zur Validierung Technischer Berechnungsgrundlagen, einen methodischen Ansatz zur Nutzung von Diagnosedaten in der Schadenbedarfsprognose, eine Ereignisprognose zu Wiedereinweisungen ins Krankenhaus sowie eine explorative Analyse der Prämienstruktur eines Krankenversicherungsbestands. Mit Ausnahme des methodischen Ansatzes steht der zugehörige Quellcode jeweils frei zur Verfügung.

5.1. Limitierungstool

Die Limitierung gehört zu den zentralen Schritten der Beitragsanpassung (vgl. Schritt 10 in Abschnitt 4.1.4): Hohe Beitragssprünge im Bestand werden durch den Einsatz von RfB-Mitteln begrenzt, wobei Wirkung und Mittelbedarf alternativer Limitierungsregeln gegeneinander abzuwägen sind. Die Arbeitsgruppe hat hierfür ein modernes Limitierungstool entwickelt, dessen Aufbau und Nutzen im Folgenden beschrieben werden. Der fachliche Kern des Limitierungstools besteht in einer mehrstufigen Simulations- und Bewertungslogik, die gezielt auf typische Fragestellungen der Beitragsanpassung ausgerichtet ist. Ausgangspunkt ist ein synthetischer Bestandsdatensatz mit Prämien vor und nach Anpassung sowie ergänzenden Tarif- und Ausscheideinformationen.

Darauf aufbauend ermöglicht das Tool erstens eine Limitierung nach definierten Regeln: Für den Gesamtbestand oder für klar abgegrenzte Teilsegmente (z. B. Tarifwerk, Tarifart, Altersgruppen, historische Belastung) können prozentuale und absolute Obergrenzen vorgegeben werden. Die Regeln werden in einer festgelegten Reihenfolge angewendet, sodass spätere Regeln frühere Ergebnisse nur verschärfen und damit eine fachlich transparente Priorisierung möglich ist. Das Ergebnis sind limitierte Anpassungen je einzelner Versicherter, die daraus abgeleitete monatliche Entlastung sowie der korrespondierende RfB-Bedarf über Einmalbeiträge.

Zweitens bildet das Tool eine Limitierung auf Personenebene für Modultarife ab. Dieser Funktionsbaustein adressiert die Praxis, dass eine Person mehrere Tarifmodule gleichzeitig führt (i. d. R. ambulant, stationär, Zahn), deren Einzelanpassungen in Summe eine zusätzliche Belastung erzeugen können. Neben den tarifbezogenen Limitierungen kann daher eine personenbezogene Gesamtlimitierung definiert werden, entweder absolut in Euro pro Monat oder relativ als Prozentsatz der aggregierten Altprämie. Damit lässt sich die Gesamtzumutbarkeit auf Personenebene systematisch steuern, ohne die tarifliche Detailstruktur aufzugeben. Die kombinierte Betrachtung aus Tarif- und Personenlogik ist insbesondere für eine differenzierte Zumutbarkeitsprüfung im Kontext von § 155 VAG relevant.

Das Limitierungstool verwendet als Dateninput exemplarisch die Ergebnisse einer Bestandssimulation. Folgende Annahmen wurden hierbei getroffen: Verwendung von Rechnungszins je Tarif, Modellierung kombinierter Ausscheidewahrscheinlichkeiten (Sterblichkeit und Storno), Umrechnung monatlicher Entlastung in kapitalisierte Einmalbeiträge über Barwertfaktoren sowie die Annahme eines gegebenen Regelsets als exogener Steuerungsinput. Für Sensitivitätsanalysen werden Zielwerte systematisch variiert, um die Wirkung auf Anzahl limitierter Versicherter, durchschnittliche Entlastung und RfB-Bedarf zu quantifizieren. Damit dient das Tool nicht nur der punktuellen Nachrechnung einzelner Maßnahmen, sondern als reproduzierbares Simulationsinstrument zur fachlichen Kalibrierung von Limitierungsstrategien unter regulatorischen und wirtschaftlichen Nebenbedingungen.

Aus softwaretechnischer Sicht zeigt das Tool den Vorteil einer modernen, modularen Architektur gegenüber historisch gewachsenen Tabellenkalkulations- und Altsystemlösungen. Während in Tabellenkalkulationen fachliche Logik, Datenaufbereitung, Berechnung und Präsentation häufig eng miteinander verschränkt sind, trennt das Limitierungstool diese Ebenen konsequent: Die Fachlogik liegt in klar abgegrenzten Services, die Benutzeroberfläche übernimmt lediglich Parametrisierung, Visualisierung und Export. Diese Struktur erhöht die Nachvollziehbarkeit, erleichtert fachliche und technische Reviews und schafft die Grundlage für automatisierte Tests. Gleichzeitig verbessert sie die Performanz, da Berechnungen auf Vektoroperationen basieren und damit auch bei größeren

Beständen und umfangreichen Szenarien deutlich robuster und schneller sind als etwa gewachsene Excel-Modelle mit zahlreichen Zellverknüpfungen. Besonders bei Sensitivitätsanalysen, segmentierten Regeln und iterativen Szenarienrechnungen zeigt sich der Mehrwert einer solchen Architektur: Ergebnisse lassen sich reproduzierbar erzeugen, Änderungen versionieren und fachliche Annahmen systematisch vergleichen. Die im Projekt vorhandenen Tests für Limitierung, Regeltracing, Sensitivität und Personen-Limitierung unterstreichen zusätzlich, dass fachliche Logik hier nicht nur implementiert, sondern auch dauerhaft abgesichert wird.

Ein weiterer wesentlicher Aspekt ist die produktivitätssteigernde Rolle von LLM-gestützter Programmierung. Sprachmodelle können die Entwicklung solcher Werkzeuge spürbar beschleunigen, etwa beim Entwurf von Funktionsgerüsten, bei der Formulierung von Testfällen, bei Refactoring-Vorschlägen, bei der Dokumentation und bei der Umsetzung von UI-Komponenten. Ihr Nutzen ist insbesondere dort hoch, wo wiederkehrende Strukturen, Standardlogik oder textuelle Artefakte erzeugt werden müssen. Der Mehrwert besteht jedoch nicht darin, die fachliche Verantwortung zu ersetzen, sondern darin, Entwicklungszeit zu reduzieren und die Iteration zwischen fachlicher Idee und technischer Umsetzung zu verkürzen. Gerade im aktuariellen Umfeld ist dabei ein kritischer Umgang mit KI-generiertem Code notwendig: Rechenlogik, Grenzfälle, Datenvalidierung und regulatorische Einordnung dürfen nicht ungeprüft übernommen werden. Für den Einsatz von KI im Aktuariat sind deshalb sowohl fachliche als auch technische Fähigkeiten erforderlich, insbesondere ein sicheres Verständnis für Beitragsanpassungslogik, Datenmodelle, Tests, Versionskontrolle und die Grenzen automatisierter Codegenerierung. Besonders wichtig ist, zwischen Assistenz und Verantwortung klar zu unterscheiden: Die KI kann Vorschläge liefern und die Entwicklung beschleunigen, aber die fachliche Plausibilisierung, die Prüfung der gesetzlichen Konformität und die finale Freigabe müssen bei der Aktuarin bzw. beim Aktuar verbleiben. Wer KI-gestützt programmiert, muss deshalb in der Lage sein, Anforderungen präzise zu formulieren, Ergebnisse kritisch zu bewerten und fachlich riskante Vereinfachungen oder stille Annahmen zuverlässig zu erkennen.

Der Quellcode des Limitierungstools steht im GitHub-Repository der Arbeitsgruppe unter einer freien Open-Source-Lizenz zur Verfügung.[20] Er darf ausprobiert, an eigene Anforderungen angepasst, weiterentwickelt und auch kommerziell eingesetzt werden.

5.2. Validierung von Technischen Berechnungsgrundlagen

Während Abschnitt 5.1 den Einsatz von LLMs bei der Entwicklung eines deterministischen Programms illustriert, ist das Sprachmodell im folgenden Praxisbeispiel Bestandteil des Verfahrens selbst: der Validierung von Technischen Berechnungsgrundlagen. Eine TB dokumentiert die Kalkulation eines Tarifs vollständig und nachvollziehbar und ist Grundlage der Prüfung durch Treuhänder und Aufsicht (vgl. Schritt 11 in Abschnitt 4.1.4). Bei jeder Beitragsanpassung ist sie fortzuschreiben: Eine Vielzahl von Werten, Tabellen, Formeln und Querverweisen muss aus der Kalkulation in das Dokument übertragen und konsistent gehalten werden; dieser Abgleich ist zeitaufwendig und fehleranfällig. Eine vollständige manuelle Prüfung eines oftmals mehrere hundert Seiten umfassenden Dokuments ist im engen Zeitplan einer Beitragsanpassung anspruchsvoll bis kaum leistbar; das Vier-Augen-Prinzip bleibt folglich häufig auf Stichproben beschränkt.

Die Arbeitsgruppe hat daher prototypisch erprobt, wie sich die TB-Validierung durch die Kombination aus regelbasierten und LLM-gestützten Prüfungen weitgehend automatisieren lässt. Als Prüfobjekt diente bewusst eine synthetische TB: ein vollständig ausgearbeitetes TB-Dokument eines fiktiven Tarifs, dessen gesamtes Zahlenwerk deterministisch aus einer programmgesteuerten Kalkulationsstrecke erzeugt wurde, die zugleich maschinenlesbare Referenzwerte bereitstellt. Das vermeidet die Vertraulichkeitsproblematik realer Unternehmensunterlagen und macht die Ergebnisse reproduzierbar; vor allem aber erlaubt es einen objektiven Wirksamkeitsnachweis: Neben der fehlerfreien Fassung wurde eine layoutidentische Zweitfassung erstellt, in die 13 dokumentierte Fehler gezielt eingebaut wurden, von Zahlendrehern in Tabellen über widersprüchliche Parameterangaben und ein falsches Normzitat bis zu einem Vorzeichenfehler im Formelsatz und einer

[20] Abrufbar unter https://github.com/DeutscheAktuarvereinigung/WorkingGroup_AI_ModernMethods_Health/tree/main/limitierungstool.

verfälscht dargestellten Kurve in einer Profilgrafik. Die Erkennungsleistung des Prüfverfahrens lässt sich damit exakt messen.

Das Prüfverfahren ist zweistufig aufgebaut: Die erste Stufe bilden regelbasierte Konsistenz- und Vollständigkeitsprüfungen, die deterministisch und exakt reproduzierbar ablaufen. Geprüft werden etwa die Auflösbarkeit interner Querverweise, der Abgleich von Inhaltsverzeichnis und Überschriften, die Vollständigkeit des Symbolverzeichnisses, rechnerische Zusammenhänge innerhalb der Ausscheideordnung, die Monotonie der Beiträge, die Konsistenz mehrfach genannter Parameter und der Rahmen-TB-Verweise sowie zitierte Rechtsnormen (etwa aus KVAV und VAG) im Abgleich mit einer kuratierten Wissensbasis. Liegen maschinenlesbare Referenzwerte aus der Kalkulation vor – etwa als Dateiexport oder als Referenzdatenbank –, tritt als besonders wirksamer Baustein der zellgenaue Abgleich der abgedruckten Tabellen hinzu.

Die zweite Stufe ergänzt LLM-gestützte Prüfungen dort, wo Sprach- und Kontextverständnis erforderlich ist. Sie betreffen die semantische Konsistenz zwischen Abschnitten und zugehörigen Anlagen, die inhaltliche Plausibilität von Normzitaten, die Rechtschreibung sowie die abgedruckten Beispielrechnungen, bei denen das Sprachmodell lediglich Formel, eingesetzte Werte und behauptetes Ergebnis extrahiert, während die Nachrechnung deterministisch erfolgt. Ein bildbasierter Kontrollschritt prüft ergänzend den gesetzten Formelapparat und erkennt damit Satzfehler, die im extrahierten Text unsichtbar bleiben. Ein weiterer bildbasierter Schritt gleicht Diagramme mit den zugehörigen Tabellen ab, indem die Kurvenverläufe vom Seitenbild abgelesen werden.

Mehrere Vorkehrungen sichern die Nachvollziehbarkeit der LLM-Komponente: Die Antworten des Sprachmodells sind über vorgegebene Schemata strukturiert („Structured Outputs“); Rechtschreibbefunde und extrahierte Zahlen müssen wörtlich im Dokument belegbar sein und werden andernfalls verworfen; Einschätzungen, die sich nicht zweifelsfrei absichern lassen, werden bei Unsicherheit als Warnung statt als Fehler ausgewiesen; alle Modellantworten werden zwischengespeichert, sodass ein Prüflauf reproduzierbar bleibt. Beim Diagrammabgleich erhält das Modell bewusst nur das Seitenbild ohne die abgedruckten Tabellenwerte („blinde“ Ablesung); der eigentliche Vergleich erfolgt anschließend deterministisch. Die Erprobung zeigte, warum das entscheidend ist: Kannte das Modell die Soll-Werte, neigte es dazu, diese schlicht zu bestätigen, und übersah dabei selbst eine um 50 % verfälschte Kurve (Verankerungseffekt). Jeder Lauf endet in einem maschinenlesbaren Prüfbericht, der je Einzelprüfung den Status und je Befund Schweregrad und Fundstelle ausweist und sich als Prüfungsevidenz archivieren lässt.

Die Erprobung bestätigt den Mehrwert der Kombination: Die regelbasierte Stufe allein identifizierte neun der 13 eingebauten Fehler; zusammen mit den LLM-gestützten Prüfungen fand ein leistungsfähiges multimodales Sprachmodell alle 13 – zwei davon, den Vorzeichenfehler im Formelsatz und die verfälschte Kurve, ausschließlich über die bildbasierten Prüfschritte. Ein kleineres, schnelleres Modell übersah diese Bildbefunde teilweise; die Modellwahl ist damit selbst Gegenstand der Validierung. Die fehlerfreie Fassung blieb ohne fälschlich gemeldete Fehler; es verblieben lediglich wenige Warnhinweise zur menschlichen Sichtung. Der Nutzen liegt dabei weniger im Ersatz der menschlichen Prüfung als in ihrer Neuausrichtung: Die maschinelle Prüfung übernimmt die vollständige Kontrolle der Detailfülle und kann nach jeder Änderung wiederholt werden, während sich die aktuarielle Durchsicht auf die fachliche Angemessenheit von Methodik und Herleitung konzentrieren kann. Fehler werden so bereits vor der Übergabe an die Treuhänderin oder den Treuhänder erkannt, und der Ansatz lässt sich mit oftmals lediglich moderatem Mehraufwand auf ganze Tarifwerke ausweiten.

Zugleich sind die Grenzen klar zu benennen. LLMs arbeiten probabilistisch; sie können Befunde übersehen oder trotz der beschriebenen Vorkehrungen unzutreffende Hinweise erzeugen. In der Erprobung stufte das Sprachmodell etwa vereinzelt korrekte Normzitate mit hoher Überzeugung als falsch ein; derartige Urteile werden deshalb nur bei offensichtlicher Themenverfehlung als Fehler, im Übrigen höchstens als prüfenswerte Warnung ausgewiesen. Ein unauffälliger Prüfbericht ist deshalb kein Nachweis der Fehlerfreiheit, sondern reduziert lediglich das Restrisiko. Daraus folgt die Notwendigkeit einer „Validierung des Validators“: Die bewusst fehlerhafte Referenzfassung dient als dauerhafter Prüfstand, an dem jede Änderung am eingesetzten Sprachmodell oder an

den Prüfregeln erneut zu messen ist. Die Verantwortung für die Korrektheit der TB verbleibt unverändert beim Fachbereich; insbesondere trägt die Verantwortliche Aktuarin oder der Verantwortliche Aktuar die fachliche Sicherstellungsverantwortung für die korrekte Prämienberechnung (vgl. Kapitel 6). Die maschinelle Prüfung bleibt damit eine unterstützende, vorbereitende Tätigkeit, die keine Entscheidungen automatisiert. Anders als beim Limitierungstool in Abschnitt 5.1, bei dem LLMs nur die Entwicklung unterstützen, ist das Sprachmodell hier allerdings Bestandteil des betriebenen Verfahrens; für die Einordnung nach AI Act und DORA gelten die in Kapitel 3 dargestellten Grundsätze. Bei Nutzung externer LLM-Dienste sind zudem Drittparteien- und Vertraulichkeitsaspekte zu klären.

Das beschriebene Validierungswerkzeug steht als dokumentiertes Jupyter-Notebook im GitHub-Repository der Arbeitsgruppe unter einer freien Open-Source-Lizenz zur Verfügung.[21] Mit ihm veröffentlicht sind die synthetische Test-TB in ihrer fehlerfreien und ihrer fehlerbehafteten Fassung, die Dokumentation der eingebauten Fehler sowie die maschinenlesbaren Referenzwerte. Damit lässt sich ein vollständiger Prüflauf nachvollziehen und die berichtete Erkennungsleistung reproduzieren. Für den Einsatz auf eigenen Dokumenten sind die Prüfregeln an die jeweilige TB-Struktur anzupassen; die bildbasierten Prüfschritte setzen zudem den Zugang zu einem multimodalen Sprachmodell voraus.

5.3. Diagnosedaten in der Schadenbedarfsprognose

5.3.1. Ausgangslage

Wie sich der Schadenbedarf entwickelt, wird meistens über Quoten prognostiziert, über das Verhältnis von beobachtetem zu rechnungsmäßigem Schaden. Daneben steht das Kopfschadenprofil, das laufend bewertet und bei Bedarf angepasst wird. Weil der rechnungsmäßige Schaden die Alterung bereits enthält, zeigt die Quote nur, was über die erwartete Alterung hinaus passiert ist. Fortgeschrieben wird damit eine einzige, stark verdichtete Restgröße.

In dieser einen Zahl stecken verschiedene Kräfte. Die medizinische Inflation treibt sie, ebenso die Menge und Intensität der Inanspruchnahme. Dazu kommt die echte Veränderung der Morbidität im Bestand und schließlich die Verschiebung der Bestandszusammensetzung durch Tod, Storno, Neugeschäft und Tarifwechsel. Die Quotenfortschreibung unterstellt, dass diese Mischung stabil bleibt; in der Praxis ist diese Annahme selten erfüllt. Hinzu kommt, dass ein vorübergehender Ausschlag und eine dauerhafte Verschiebung in der Quote gleich aussehen. Prozessual wird dieses Thema auch in Abschnitt 4.1.2 aufgegriffen.

5.3.2. Der Mehrwert der Aufteilung

Der hier vorgestellte Ansatz zerlegt die Quote in ihre Bestandteile und macht damit sichtbar, woraus sich der beobachtete Trend zusammensetzt. Das verbessert das Verständnis der Bestandsentwicklung und damit die Prognosequalität. Jeder Bestandteil lässt sich anschließend sachgerecht fortschreiben: die strukturelle Morbidität demografisch und epidemiologisch, die medizinische Inflation anhand geeigneter interner und externer Quellen, ein lediglich vorübergehender Ausschlag dagegen gar nicht. Die Morbidität wirkt darüber hinaus als Frühwarnindikator: Steigt die Prävalenz chronischer Erkrankungen, zeichnet sich ein Anstieg ab, bevor er in der Quote sichtbar wird. Die Trennung von dauerhaften und einmaligen Effekten liefert zugleich eine belastbare Begründung gegenüber der Treuhänderin oder dem Treuhänder und verhindert, dass ein einmaliger Ausreißer dauerhaft in den Beitrag einfließt.

[21] Abrufbar unter https://github.com/DeutscheAktuarvereinigung/WorkingGroup_AI_ModernMethods_Health/tree/main/tb-validierung.

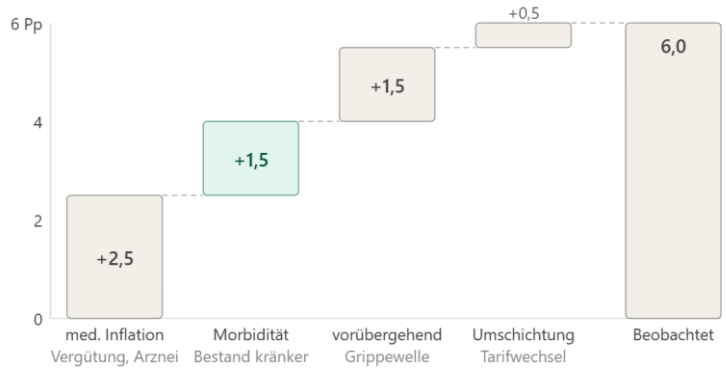


Abbildung 2: Schematische Zerlegung eines beobachteten Schadentrends in seine Bestandteile (illustratives Beispiel; Angaben in Prozentpunkten).

5.3.3. Ein möglicher Ansatz

Der Morbi-RSA der GKV liefert ein gutes Vorbild für die Berechnung der Morbidität: Diagnosen werden zu Morbiditätsgruppen zusammengefasst, die nach Schweregrad gestuft sind, und über kostenbasierte Gewichte zu einem Index verdichtet, der sich als gewichtete Summe der Gruppen lesen lässt. Dieses Grundprinzip lässt sich übernehmen, der Weg zu einer belastbaren Trendaussage in der PKV verlangt aber einige zusätzliche Schritte. Diese sind unter anderem:

- **Auf Einzelaltern messen.** Die altersspezifischen Werte werden mit einer fest gewählten Altersstruktur zu einem Jahreswert verdichtet. So verschiebt sich dieser Wert nicht schon deshalb, weil der Bestand insgesamt altert, sondern nur, wenn die Morbidität bei gegebenem Alter zunimmt.
- **Krankheiten aus mehreren Quellen ableiten**. Nicht nur aus Diagnosen, sondern auch aus Arzneimitteln und Behandlungen. In der PKV ist die ambulante Codierung zu schwach, um sich allein darauf zu verlassen. Dazu klare Aufgreifregeln: nur gesicherte Diagnosen, Mehrfachbestätigung, chronische Erkrankungen einmal erkannt fortschreiben.
- **Einen Score mit eingefrorenen Gewichten bilden.** Die Schweregewichte werden einmal auf einer Referenzperiode geschätzt und dann konstant gehalten. So fließen Kosten in die Skala ein, aber nicht mehr in ihre Bewegung. Der Score verändert sich nur noch mit der Krankheitszusammensetzung. Die Gruppenstruktur trägt man stets mit, sonst lässt sich eine Veränderung später nicht erklären.
- **Den stehenden Bestand von seiner Bewegung trennen.** Die echte Verschlechterung zeigt sich an den Verbleibern von Jahr zu Jahr, verglichen mit dem, was ihr Älterwerden allein erwarten ließe. Tarifwechsel nach § 204 VVG, Stornoselektion und Neugeschäft dagegen verschieben den Score, ohne dass jemand kränker wird, und gehören separat ausgewiesen.
- **Den Score auf den Schaden normieren.** Der Score erfasst nur den Block, den die Morbiditätsgruppen abdecken. Um den Beitrag zum Gesamttrend zu erhalten, gewichtet man die Veränderung mit dem Anteil, den der Block am Schaden hat.

5.3.4. Interpretierbare und Black-Box-Modelle

Am einfachsten setzt man die Gewichte für den Morbiditätsscore direkt, etwa als mittlere Leistung je Gruppe auf der Referenzperiode oder als fachlich festgelegte Schweregrade. Das ist maximal transparent, aber grob: Wer in mehreren Gruppen steckt, wird doppelt gezählt, und Alters- wie Begleiterkrankungseffekte vermischen sich, weil nichts gegeneinander kontrolliert wird.

Die nächste Stufe behebt genau das durch eine Regression der Leistung auf die Gruppenindikatoren, wie sie auch der Morbi-RSA nutzt. Ein Generalisiertes Lineares Modell (GLM), das für die schiefe und nulllastige Leistungsgröße sinnvollerweise mit einer Tweedie- oder Gamma-Verteilung angesetzt wird, schätzt die Gewichte gemeinsam und liefert damit den partiellen Beitrag jeder Gruppe – bereinigt um Alter und um das gleichzeitige Vorliegen anderer Erkrankungen. Ein Generalisiertes Additives Modell (GAM) erlaubt zusätzlich einen glatten, nicht-linearen Alterseffekt, ohne die Interpretierbarkeit aufzugeben: Jede Gruppe behält ihren nachvollziehbaren Koeffizienten. Diese Stufe ist der natürliche Kandidat für die produktive Bestimmung der Gewichte, weil sie Trennschärfe und Nachvollziehbarkeit verbindet.

Die dritte Stufe sind Black-Box-Verfahren wie Gradient Boosting. Sie sind am treffsichersten, für die produktive Gewichtsbestimmung aber ungeeignet wegen Nachvollziehbarkeit, Stabilität und regulatorischer Anforderungen. Ihr Platz ist ein anderer und durchaus wertvoll: als Prüfinstrument für die Regressionsstufe. Sie zeigen, wo die additiv geschätzten Gewichte danebenliegen, weil Multimorbidität nicht additiv wirkt – zwei chronische Erkrankungen zusammen kosten i. d. R. mehr als ihre Summe –, und sie wirken als Suchscheinwerfer für unplausible Muster, die meist kein medizinisches Signal sind, sondern ein Datenartefakt, etwa die Codiereffekte, die das Morbiditätssignal verzerren. Was die Black Box findet, fließt nicht in eine eigene Black-Box-Skala, sondern zurück ins GLM. So bleibt das produktive Modell interpretierbar und gewinnt zugleich an Treffsicherheit. Eine Prognose des Scores je Versicherten wird dagegen bewusst nicht verfolgt: Für die aggregierte Sicht ist sie unnötig und als Individualscoring regulatorisch heikel.

5.3.5. Herausforderungen und Fazit

In der Umsetzung ist eine Reihe von Details zu beachten; herausgegriffen sei die Erfassungs- oder Codierreife.

Wird die Codierung über die Jahre vollständiger, steigt die gemessene Prävalenz, ohne dass die Morbidität tatsächlich zunimmt. Dass dieser Effekt praktisch relevant ist, zeigt die GKV, die ihm mit einer eigenen Manipulationsbremse begegnet. In der PKV können erfassungsstabile Ankererkrankungen diese Kontrollfunktion übernehmen.

Zusammenfassend schreibt die Quotenfortschreibung einen vermischten Trend fort und unterstellt dabei, dass dessen Zusammensetzung unverändert bleibt. Die genauere Auswertung der Leistungs- und Diagnosedaten löst diese Vermischung auf: Sie zeigt, woraus sich der Trend zusammensetzt, und erlaubt es, jeden Bestandteil sachgerecht fortzuschreiben. Das Ergebnis ist eine genauere und besser begründbare Prognose.

5.4. Fallprognose: Wiedereinweisung ins Krankenhaus

Die in Anhang A beschriebenen Grundlagen des maschinellen Lernens werden im folgenden Anwendungsfall exemplarisch angewendet.

5.4.1. Schätzung der Wahrscheinlichkeit für eine Wiedereinweisung

In diesem Prognose-Notebook wird mittels Methoden des maschinellen Lernens die Wiedereinweisungswahrscheinlichkeit in ein Krankenhaus unter Verwendung von Einzeldiagnosen und zahlreichen weiteren Merkmalen vorhergesagt.

Ziel ist es, zu zeigen, wie mit den beschriebenen Verfahren eine methodisch saubere Ereignisprognose in Python erstellt werden kann und welche Tools benötigt werden bzw. für diesen Anwendungsfall besonders geeignet und erfolgreich sind.

Aus Gründen der Übersichtlichkeit wird von den Gradient-Boosting-Verfahren lediglich CatBoost betrachtet, da dieses Verfahren im vorliegenden Anwendungsfall besonders gute Ergebnisse erzielt.

Ein Vergleich mit anderen Gradient-Boosting-Verfahren wie LightGBM und XGBoost findet sich zum Beispiel im Jupyter-Notebook „Binary Classification Use Case: Credit Scoring“, das als ADS Use Case auf dem GitHub-Account der Deutschen Aktuarvereinigung zur Verfügung steht.[22]

5.4.2. Patientenhistorie als Ausgangspunkt

Als Datenbasis dient der öffentlich verfügbare Datensatz „Diabetes 130-US Hospitals for Years 1999-2008“ aus dem UCI Machine Learning Repository,[23] der zusammen mit dem Artikel von Strack et al. (2014) veröffentlicht wurde.[24]

Der zusammen mit dem oben genannten Artikel veröffentlichte Datensatz enthält ca. 100.000 US-amerikanische Krankenhausfälle mit 47 Merkmalen. Zu den enthaltenen Informationen zählen unter anderem drei ICD-Diagnosen sowie Angaben darüber, ob eine Wiedereinweisung innerhalb kurzer Zeit, zu einem späteren Zeitpunkt oder gar nicht erfolgte.

Der Datensatz stellt eine außergewöhnlich umfangreiche und frei verfügbare Grundlage nicht nur für Untersuchungen zu Krankenhauswiedereinweisungen dar. Vergleichbare öffentlich zugängliche, branchenspezifische Datensätze aus Deutschland sind nicht bekannt.

5.4.3. Ergebnis: Vergleich von Modellansätzen

Das Analyse-Notebook „hospital-readmission“ steht im GitHub-Repository der Arbeitsgruppe als ausführlich dokumentierter HTML-Analysebericht und als ausführbares Jupyter-Notebook (.ipynb) zur Verfügung.[25]

Es lässt sich in die eigene Entwicklungsumgebung übernehmen und so anpassen, dass sich die technischen Funktionalitäten auf andere Datensätze und Fragestellungen übertragen lassen, etwa auf eine Stornoprognose.

Erstellt wurde das Notebook mit Unterstützung eines Sprachmodells. Das zeigt, wie effizient sich mit Werkzeugen wie CatBoost und AutoGluon (siehe Abschnitt A.3.1) inzwischen auch ohne umfangreiche Programmierkenntnisse belastbare Ereignisprognosen erstellen lassen.

5.5. Analyse der Prämienstruktur eines Krankenversicherungsbestands

5.5.1. Zielsetzung: Bestandsstrukturen transparent machen

In diesem Analyse-Notebook wird die Prämienstruktur eines Krankenversicherungsportfolios in einem R-Notebook explorativ analysiert, und die dabei aufgedeckten Zusammenhänge grafisch visualisiert. Darüber hinaus werden eine Reihe von eher klassischen statistischen Modellen getestet, die diese Prämienstruktur nachbilden sollen, indem sie die wesentlichen Einflussgrößen auf die Prämienstruktur identifizieren und modellieren.

Ziel ist es zu zeigen, wie mit Open-Source-Analytics ein Datenbestand untersucht werden kann, der neben einer longitudinalen Komponente (dem Alter) noch mehrere diskrete Komponenten enthält. Ein solcher Datensatz tritt in der Krankenversicherung häufig auf, zum Beispiel bei Untersuchungen der eingetretenen Schäden.

---

[22] Abrufbar unter https://github.com/DeutscheAktuarvereinigung/ADS_Use_Cases/blob/main/binary-classification-credit-scoring.ipynb.

[23] Clore, J., Cios, K., DeShazo, J., Strack, B. (2014): *Diabetes 130-US Hospitals for Years 1999-2008*, UCI Machine Learning Repository, abrufbar unter https://doi.org/10.24432/C5230J

[24] Strack, B. et al. (2014): *Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records*, BioMed Research International, Band 2014, Artikel-ID 781670, abrufbar unter https://doi.org/10.1155/2014/781670

[25] Abrufbar unter https://github.com/DeutscheAktuarvereinigung/WorkingGroup_AI_ModernMethods_Health/tree/main/fallprognose-wiedereinweisung.

Das Beispiel illustriert auch, auf welche Grenzen diese Untersuchungen stoßen, wenn die Struktur der zugrunde liegenden Daten durch nicht explizit dargestellte Eigenschaften beeinflusst wird, wie etwa sich ändernde Produktlinien, oder Höhe und Anlass von Risikozuschlägen.

5.5.2. Datenbasis

Als Datenbasis dient ein öffentlich verfügbarer Datensatz zur spanischen Krankenversicherung, der zusammen mit dem zugehörigen Artikel von Lledó et al. (2026) veröffentlicht wurde. [26]

Es handelt sich um einen anonymisierten Vertragsbestand eines spanischen Krankenversicherers mit verschiedenen Produktlinien (Standard, Premium, International, Dental) sowie Angaben zu Alter, Geschlecht, Vertriebsweg, Vertragsart und Vertragsdauer einschließlich Neugeschäft und Storno. Der Bestand umfasst rund 225.000 Datensätze zu über 70.000 versicherten Personen aus den Jahren 2017 bis 2019. Ergänzt wurde er um geografische und sozioökonomische Merkmale aus öffentlich zugänglichen Quellen, unter anderem des spanischen Statistikinstituts; diese sind in der vorliegenden Analyse nicht einbezogen.

Vergleichbare öffentlich zugängliche, branchenspezifische Vertragsdaten aus Deutschland sind nicht bekannt. Die verwendeten Methoden lassen sich jedoch ohne Weiteres auf unternehmenseigene Datensätze übertragen.

5.5.3. Ergebnisse: Auswirkung struktureller Bestandsmerkmale

Das Analyse-Notebook „praemienstruktur“ wurde mit der neben Python ebenfalls weit verbreiteten und für solche Analysen gut geeigneten Statistik-Programmiersprache R erstellt. Es steht im GitHub-Repository der Arbeitsgruppe als ausführlich dokumentierter HTML-Analysebericht und als ausführbarer Code (.Rmd) zur Verfügung.[27] Letzterer kann kopiert, geändert und auf andere, gegebenenfalls unternehmenseigene Datensätze angewandt werden.

Das Notebook setzt keine Vorkenntnisse im Bereich Machine Learning voraus und eignet sich daher besonders als niedrigschwelliger Einstieg in die Welt der Open-Source-Analytik.

Auf dem GitHub-Account der Deutschen Aktuarvereinigung sind zudem mehrere Notebooks zu modellagnostischen Erklärbarkeitsmethoden sowie weitere interessante Beiträge anderer DAV-Arbeitsgruppen verfügbar.[28]

[26] Lledó, J., Espinosa, P., Pérez, V. (2026): *A dataset for health insurance analysis: Integrating individual and area-based contextual variables*, Scientific Data 13, Artikel 54, abrufbar unter https://doi.org/10.1038/s41597-025-06372-z

[27] Abrufbar unter https://github.com/DeutscheAktuarvereinigung/WorkingGroup_AI_ModernMethods_Health/tree/main/praemienstruktur.

[28] Abrufbar unter https://github.com/DeutscheAktuarvereinigung/.

6. Rolle und Verantwortung der Aktuarinnen und Aktuare

Dieses Kapitel beschreibt, wie Aktuarinnen und Aktuare in der privaten Krankenversicherung den Einsatz von KI und modernen Methoden fachlich beurteilen und verantworten können. Darüber hinaus wird in Abschnitt 6.4 die Position der DAV zur regulatorischen Einordnung klassischer statistischer Verfahren dargestellt.

6.1. Einordnung und Bewertung des Einsatzes von KI und modernen Methoden in der privaten Krankenversicherung

Die vorhergehenden Kapitel haben aufgezeigt, dass es eine Vielzahl von Anwendungsfällen für den Einsatz von KI und modernen Methoden in der privaten Krankenversicherung gibt. Diese weisen – richtig eingesetzt – einen erkennbaren Nutzen auf: Sie

- beschleunigen einzelne Arbeitsschritte, Prozessabschnitte oder sogar vollständige Prozesse,
- ermöglichen Analysen und leiten daraus Erkenntnisse ab, die mit klassischen Methoden teilweise nur mit sehr hohem Aufwand gewonnen werden könnten, oder
- bieten eine deutliche Unterstützung bei Dokumentationen oder empfängergerechten Darstellungen.

Zugleich wird aufgezeigt, dass die ausgeprägte Regulatorik der privaten Krankenversicherung zwar einen wichtigen Rahmen darstellt, den es beim Einsatz von KI und modernen Methoden zu berücksichtigen gilt. Dieser Rahmen steht aber ihrem Einsatz keineswegs entgegen. Vielmehr bieten die für derartige Anwendungen geschaffenen Gesetzesgrundlagen des AI Acts und DORA Leitplanken, die auch für die private Krankenversicherung einen weiten Handlungsspielraum aufzeigen, vielfältigste Anwendungsmöglichkeiten für KI und moderne Methoden im aktuariellen Umfeld ermöglichen und anwendungsorientiert regeln. Bestehende regulatorische Unsicherheiten in Bezug auf die Einordnung bestimmter, auch bereits lange im Einsatz befindlicher klassischer Methoden, führen nicht dazu, dass diese oder auch neue Methoden generell nicht angewendet werden können. Ihre Anwendung ist grundsätzlich möglich, sofern entsprechende Vorkehrungen getroffen werden.

6.2. Verantwortung der Aktuarinnen und Aktuare beim Einsatz von KI

Kapitel 3 beschreibt die regulatorischen und organisatorischen Anforderungen an KI-Systeme, moderne Methoden und IKT-Anwendungen, Daten und Modelle. Dort werden unter anderem die Einordnung nach dem AI Act, die Anforderungen aus DORA sowie die Vorgaben zu Daten, Dokumentation, Versionierung, Tests, menschlicher Aufsicht, Änderungen und Vorfällen behandelt.

6.2.1. Einordnung von KI und modernen Methoden und abzuleitende Maßnahmen

Aus den dargestellten gesetzlichen Grundlagen leiten sich konkrete Anforderungen ab, die Aktuarinnen und Aktuare im Zusammenhang mit dem Einsatz von KI und modernen Methoden berücksichtigen sollten. Nachfolgende Übersicht fasst wesentliche Fragestellungen zusammen, die Aktuarinnen und Aktuare beim Einsatz derartiger Anwendungen beleuchten sollten:

Fragestellung	Kriterien	Maßnahmen
Liegt ein KI-System im Sinne des AI Acts vor?	Ist die Anwendung ein KI-System (z. B. ML-Modell, LLM-gestützte Fachanwendung)?	Prüfen und dokumentieren.
Wird die Anwendung in der Rolle als Anbieter oder Betreiber eingesetzt?	Wer legt Zweck und fachliche Aufgabe des Systems fest? Wer bestimmt Daten, Merkmale, Modellwahl und Validierungsanforderungen? Wer gibt das System frei und entscheidet über Änderungen? Je stärker das Unternehmen (bzw. das Aktuariat) Zweck, Datenbasis und Funktionsweise prägt, desto eher liegt eine Anbieterrolle vor – mit entsprechend strengeren AI-Act-Pflichten.	Prüfen und auf Einhaltung der Anforderungen hinwirken bzw. Mitwirkung im erforderlichen Maße, u. a. Sicherstellung einer ausreichenden Befähigung sowie von weiteren Anforderungen, insbesondere bei Hochrisiko-Anwendungen. Juristisch ist das Unternehmen Anbieter/Betreiber, nicht die Aktuarin bzw. der Aktuar selbst.
Handelt es sich um eine Anwendung, die den verbotenen Praktiken zuzuordnen ist?	Führt die Anwendung zu einer Bewertung des Sozialverhaltens natürlicher Personen, die in einem anderen Zusammenhang (hier im Rahmen der privaten Krankenversicherung) verwendet wird und dort zu einer schweren Benachteiligung führt?	Prüfen, ob Kriterien der verbotenen Praktiken auf die Anwendung zutreffen. Sofern dies der Fall ist, darf die Anwendung in der geplanten Form nicht in Betrieb genommen werden.
Handelt es sich um eine Hochrisiko-Anwendung?	Dient das System zur Risikobewertung oder Preisbildung in der privaten Krankenversicherung oder wird ein Profiling anhand von Daten vorgenommen? Handelt es sich um einen Einsatz bei vorbereitenden Aufgaben, hat das Ergebnis keinen wesentlichen Einfluss auf Entscheidungen und findet kein Profiling statt?	Bei Einstufung als Hochrisiko-System: Hinwirken auf die Etablierung eines adäquaten Risikomanagement- und Validierungssystems im Rahmen von Konzeption, Umsetzung, Einsatz, Änderung; Grundrechte-Folgenabschätzung, Etablierung von menschlichen Kontrollen, etc.
Handelt es sich um eine Anwendung, die besonderen Transparenzanforderungen unterliegt?	Müssen die Nutzer einer Anwendung darauf hingewiesen werden, dass sie mit einer KI interagieren bzw. ist sichergestellt, dass die Funktionen nicht manipulativ oder irreführend sind?	Prüfen und ggf. geeignete Maßnahmen ergreifen (z. B. durch entsprechende Hinweistexte).
Welche DORA-Kritikalität besitzt die Anwendung?	Fällt die Anwendung unter die DORA-Governance als IKT-Anwendung (mit kritischer/wichtiger Funktion)?	Bei kritischen Anwendungen: Neu entwickelte oder geänderte Tools sollten einem geregelten Änderungs-, Test- und Freigabeprozess unterliegen. Auch fachliche Tests sind erforderlich.

Welche Anforderungen sind im Zusammenhang mit der Nutzung der verwendeten Daten zu beachten (Trainingsdaten, Analysedaten)?	Relevanz, Repräsentativität, Vollständigkeit, Datenqualität, Verzerrungsfreiheit, Zweckbestimmung, Schutzanforderungen (z. B. bei Gesundheitsdaten)	Sicherstellen der Datenqualität, Datennutzbarkeit (z. B. Einsatz der Daten für den Zweck zulässig, Einholen von Einwilligungserklärungen, o. Ä.), adäquate Datenaufbereitung (z. B. Pseudonymisierung bei Gesundheitsdaten), Vermeidung von Datenverzerrungen
Ist die Anwendung erklärbar?	Welches Maß an Erklärbarkeit ist in technischer und inhaltlicher Hinsicht erforderlich? Sind Ergebnisse nachvollziehbar bzw. prüfbar und reproduzierbar?	Vorab definieren, was in welchem Umfang erklärbar sein soll, Methoden festlegen und dokumentieren.

Tabelle 2: Fragestellungen zur Einordnung von Anwendungen, die im Aktuariat eingesetzt werden.

6.2.2. Fachliche Beurteilung wesentlicher Anwendungen

Für Aktuarinnen und Aktuare stellt sich insbesondere beim Einsatz von KI-Anwendungen und modernen Methoden die Frage, wie sie eine konkrete Anwendung fachlich beurteilen und unter welchen Voraussetzungen deren Ergebnisse in einem aktuariellen Prozess verwendet werden können. Ihre Aufgabe besteht vor allem darin, die fachliche Eignung der verwendeten Daten, Methoden und Ergebnisse zu bewerten.

Die Tiefe der fachlichen Beurteilung richtet sich nach der möglichen Auswirkung der Anwendung. Bei allgemeinen Arbeitsmitteln ohne unmittelbaren Einfluss auf produktive Berechnungen oder fachliche Festlegungen genügen regelmäßig die üblichen Sorgfaltspflichten. Dies betrifft beispielsweise Recherchen, sprachliche Überarbeitungen, Präsentationsentwürfe und explorative Analysen. Die Aktuarin bzw. der Aktuar prüft das Arbeitsergebnis vor seiner weiteren Verwendung und beachtet die unternehmensweiten Vorgaben zum Umgang mit Daten.

Einen höheren Prüfungsbedarf haben fachliche Hilfsmittel, die Hinweise, Vorschläge oder Entwürfe für eine anschließende aktuarielle Beurteilung liefern. Hierzu gehören etwa Anomalieerkennungen, Vergleichsmodelle, LLM-gestützte Textentwürfe und automatisierte Konsistenzprüfungen. In diesen Fällen sollte die Aktuarin bzw. der Aktuar beurteilen, ob das Hilfsmittel für den vorgesehenen Zweck geeignet ist und welche Prüfung vor der Verwendung seiner Ergebnisse erforderlich ist. Abhängig von der möglichen Fehlerwirkung können Stichproben, repräsentative Testfälle oder Vergleichsrechnungen ausreichen.

Eine vertiefte fachliche Beurteilung ist erforderlich, wenn eine Anwendung Prämien, Rückstellungen oder Rechnungsgrundlagen unmittelbar berechnet oder deren Festlegung wesentlich beeinflusst. Gleiches gilt, wenn sie Entscheidungen über einzelne Versicherte erheblich beeinflusst, bei Fehlern wesentliche Auswirkungen auf die Kalkulation haben kann, als Hochrisiko-KI-System eingestuft ist oder nach der DORA-Governance des Unternehmens eine kritische oder wichtige Funktion unterstützt. Die Einstufung als Hochrisiko-System wird in der aktuariellen Arbeit in der privaten Krankenversicherung in vielen Fällen gegeben sein.

Diese Einordnung bezieht sich auf die produktiv verwendete Fachanwendung. Der Einsatz von KI während der Entwicklung ist davon getrennt zu betrachten. Wird ein Sprachmodell beispielsweise zur Erstellung von Programmcode verwendet, ist es zunächst ein Entwicklungswerkzeug. Die erstellte Fachanwendung wird anhand ihrer eigenen Ergebniswirkung beurteilt. Eine mit KI-Unterstützung programmierte Kalkulationsanwendung kann daher eine wesentliche Anwendung sein, auch wenn sie im produktiven Betrieb deterministisch arbeitet und selbst kein KI-System im Sinne des AI Acts darstellt.

Die fachliche Beurteilung einer wesentlichen Anwendung sollte folgende Fragen beantworten. Das Ergebnis sollte geeignet dokumentiert werden. Hat eine Anwendung wesentlichen Einfluss auf die Herleitung oder Festlegung von Rechnungsgrundlagen, sollte ihre methodische Bedeutung in den Technischen Berechnungsgrundlagen nachvollziehbar sein. Die Darstellung sollte erkennen lassen, welchen Beitrag die Anwendung leistet, wie ihre Ergebnisse geprüft wurden und welche fachliche Entscheidung auf dieser Grundlage getroffen wurde. Bei allgemeinen Arbeitsmitteln und Hilfsmitteln ohne Einfluss auf Methodik, Parameter oder Ergebnisse kann eine Darstellung in den Technischen Berechnungsgrundlagen entbehrlich sein. Maßgeblich ist, ob die Darstellung für die vollständige und nachvollziehbare Dokumentation der Kalkulation erforderlich ist.

Für welchen Zweck ist die Anwendung geeignet?

Der fachliche Verwendungszweck ist so zu beschreiben, dass der geprüfte Einsatzbereich eindeutig erkennbar ist. Aus der Beschreibung sollte hervorgehen, welche Aufgabe die Anwendung erfüllt, welche Ergebnisse sie erzeugt und wie diese Ergebnisse in den weiteren Prozess eingehen. Ebenso sollte erkennbar sein, welche Entscheidungen weiterhin durch einen Menschen getroffen werden und welche Nutzungen durch die fachliche Prüfung nicht abgedeckt sind.

Die regulatorische Einordnung und die organisatorischen Zuständigkeiten richten sich nach Kapitel 3 und den Prozessen des Unternehmens. Die Aktuarin bzw. der Aktuar stellt hierfür die erforderlichen fachlichen Informationen bereit, insbesondere zum geplanten Einsatz, zur Ergebniswirkung und zu den möglichen Folgen fehlerhafter Ergebnisse.

Sind die verwendeten Daten und die gewählte Methode für diesen Zweck angemessen?

Es ist zu beurteilen, ob die verwendeten Daten und die gewählte Methode für den vorgesehenen Zweck geeignet sind. Bei den Daten sind insbesondere der sachliche Bezug zum Anwendungsfall, die Abdeckung relevanter Zeiträume und Teilbestände sowie bekannte Mängel und Strukturbrüche zu betrachten. Änderungen der Erfassungspraxis, fehlende Werte, Ausreißer und kleine Bestände können die Ergebnisse wesentlich beeinflussen und sind daher in die Beurteilung einzubeziehen.

Bei der Methode ist zu prüfen, ob ihre Annahmen zum Anwendungsfall passen und ob ihre Ergebnisse hinreichend stabil und übertragbar sind. Dazu gehört auch die Frage, ob bestimmte Tarife oder Teilbestände außerhalb des belastbar untersuchten Bereichs liegen. Bei datengetriebenen Modellen ist außerdem zu beurteilen, ob mögliche Verzerrungen der Daten zu sachlich unangemessenen Ergebnissen führen können.

Datenschutzrechtliche, technische und informationssicherheitsbezogene Prüfungen erfolgen nach den dafür vorgesehenen Unternehmensprozessen.

Belegt die Validierung die fachliche Eignung?

Die Validierung soll belegen, dass die Anwendung für ihren vorgesehenen Zweck fachlich geeignet ist. Art und Umfang der Prüfung richten sich nach der möglichen Fehlerwirkung. Bei einer unmittelbar kalkulationswirksamen Anwendung wird regelmäßig ein belastbarer Rechenfallkatalog erforderlich sein. Bei einem ergänzenden Plausibilisierungsverfahren können repräsentative Stichproben und Vergleichsrechnungen genügen.

Für wesentliche Anwendungen kommen insbesondere folgende Prüfungen in Betracht:

- Vergleich mit einem geeigneten Referenzverfahren,
- Test typischer Rechenfälle,
- Prüfung fachlich relevanter Grenzfälle,
- Kontrolle festgelegter Toleranzgrenzen,
- Sensitivitätsanalysen für wesentliche Eingangsgrößen,
- Tests auf relevanten Tarifen und Teilbeständen,

- Prüfung anhand von Daten, die nicht zur Modellentwicklung verwendet wurden,
- Regressionstests bei Änderungen bestehender Anwendungen.

Für die fachliche Aussagekraft der Validierung muss eindeutig erkennbar sein, welcher Stand geprüft wurde. Aus den vorhandenen Nachweisen sollte hervorgehen, welche Modell- oder Programmversion, welcher Datenstand und welche Parameter der Prüfung zugrunde lagen. Ebenso sollten die durchgeführten Tests, die festgestellten Abweichungen und bekannte Einschränkungen nachvollziehbar sein.

Wie dürfen die Ergebnisse verwendet werden?

Die fachliche Beurteilung sollte in einer klaren Aussage dazu münden, wie die Ergebnisse der Anwendung verwendet werden dürfen. Je nach Anwendungsfall kann vorgesehen werden, dass Ergebnisse innerhalb festgelegter Toleranzen unmittelbar übernommen werden dürfen, vor der Übernahme einer zusätzlichen Plausibilitätsprüfung unterliegen oder ausschließlich als Vergleichs- und Warninformation dienen.

Für bestimmte Ausgaben kann eine verpflichtende Einzelfallprüfung vorgesehen werden. Ebenso kann festgelegt werden, dass bei definierten Auffälligkeiten auf ein Referenzverfahren zurückzugreifen oder die Verwendung der Anwendung auszusetzen ist.

Bei unterstützenden KI-Anwendungen sollte erkennbar bleiben, welchen Beitrag das System leistet und an welcher Stelle die eigenständige fachliche Entscheidung erfolgt. Eine allgemeine Aussage, dass die Ausgaben des Systems plausibel erscheinen, ist als Ergebnis der fachlichen Prüfung nicht ausreichend.

Wann ist eine erneute Prüfung erforderlich?

Die fachliche Eignung einer Anwendung kann sich im Zeitablauf verändern. Die Aktuarin bzw. der Aktuar sollte daher festlegen, anhand welcher Beobachtungen erkannt werden kann, dass die ursprüngliche Beurteilung möglicherweise nicht mehr trägt. Hierzu können Abweichungen von Referenzrechnungen, eine nachlassende Prognosegüte, eine steigende Zahl von Fehlalarmen, Veränderungen der Eingabedaten oder Auffälligkeiten auf einzelnen Teilbeständen gehören.

Eine erneute fachliche Prüfung ist erforderlich, wenn Änderungen die Ergebnisse oder den geprüften Verwendungszweck beeinflussen können. Dies betrifft insbesondere Änderungen der Berechnungslogik oder des Modells, wesentliche Anpassungen von Parametern und Datenquellen, die Übertragung auf andere Tarife oder Bestände sowie Erweiterungen des Verwendungszwecks. Auch Aktualisierungen externer Modelle oder KI-Dienste können eine erneute Prüfung auslösen, wenn sich hierdurch die erzeugten Ergebnisse verändern können.

6.3. Besondere Rolle der Verantwortlichen Aktuarin oder des Verantwortlichen Aktuars beim Einsatz von KI oder modernen Methoden

Die Verantwortliche Aktuarin oder der Verantwortliche Aktuar ist nicht allgemein für die eingesetzten Anwendungen oder deren technischen Betrieb verantwortlich. In der substitutiven Krankenversicherung hat sie oder er gemäß § 156 Absatz 2 Nummer 1 VAG jedoch sicherzustellen, dass bei der Berechnung der Prämien und versicherungstechnischen Rückstellungen die vorgeschriebenen versicherungsmathematischen Methoden eingehalten werden.

Ihre oder seine Einbindung ist daher insbesondere erforderlich, wenn eine Anwendung Prämien oder versicherungstechnische Rückstellungen berechnet, wesentliche Rechnungsgrundlagen bestimmt oder deren Festlegung maßgeblich beeinflusst. Gleiches gilt, wenn kalkulationsrelevante Daten wesentlich verändert oder selektiert werden sowie wenn Fehler der Anwendung erhebliche Auswirkungen auf die Kalkulation haben können.

Die Verantwortliche Aktuarin oder der Verantwortliche Aktuar muss die fachliche Prüfung nicht selbst durchführen. Sie oder er sollte sich auf die im Unternehmen eingerichteten Prozesse, fach-

lich geeignete Personen und dokumentierte Kontrollen stützen. Die vorhandenen Nachweise müssen ihr oder ihm eine Beurteilung ermöglichen, ob der Verwendungszweck hinreichend bestimmt wurde, Daten und Methoden fachlich geeignet sind und die Validierung die wesentlichen Fehlermöglichkeiten abdeckt.

Darüber hinaus sollte nachvollziehbar sein, wie die Ergebnisse verwendet werden dürfen, welche Änderungen eine erneute Prüfung auslösen und wie wesentliche Fehler behandelt werden. Auf diese Weise kann die Verantwortliche Aktuarin oder der Verantwortliche Aktuar ihrer bzw. seiner Sicherstellungsverantwortung nachkommen, ohne jeden technischen Entwicklungsschritt selbst zu prüfen.

Ob eine unmittelbare Einbindung der Verantwortlichen Aktuarin oder des Verantwortlichen Aktuars erforderlich ist, richtet sich nach dem konkreten Verwendungszweck und der möglichen Auswirkung der Anwendung. Bei allgemeinen Arbeitsmitteln und fachlichen Hilfsmitteln ohne wesentlichen Einfluss auf Prämien, versicherungstechnische Rückstellungen oder Rechnungsgrundlagen wird sie regelmäßig nicht erforderlich sein.

6.4. Position der DAV zur Einordnung statistischer Verfahren als KI

Derzeit findet ein intensiver Diskurs auf europäischer Ebene zur Frage statt, welche Verfahren und Tools, die teilweise schon sehr lange in der aktuariellen Arbeit Anwendung finden, unter die Regelungen des AI Acts fallen. Erheblichen Auslegungsspielräumen sind dabei insbesondere klassische statistische Verfahren unterworfen, wie beispielsweise lineare und logistische Regression als Spezialfälle von GLMs sowie GLMs als Spezialfälle von GAMs. Sie gehören in der Versicherungsmathematik seit drei Jahrzehnten zum methodischen Standardrepertoire und haben sich z. B. in der Risikoprüfung nachhaltig bewährt. Eine methodische Einordnung dieser Verfahren gibt Anhang A (Abschnitt A.2.2).

Die DAV bringt sich in die Diskussion aktiv ein. Hier sei insbesondere auf die Stellungnahme „DAV plädiert für eine rechtliche Klarstellung im AI Act zur Abgrenzung klassischer statistischer Verfahren in der KI-Definition“ vom 08.06.2026 verwiesen.[29]

Für die Bewertung, ob ein System als KI-System einzustufen ist, müssen insbesondere drei Kriterien untersucht werden: Prüfung, inwiefern

- ein autonomer Betrieb stattfindet,
- eine Selbstanpassung nach Inbetriebnahme vorgesehen ist,
- der Lösungsraum begrenzt sowie Ergebnisse erklärbar und reproduzierbar sind.

Eine pauschale Subsumtion klassischer statistischer Verfahren, welche jahrzehntelang zuverlässig in der Versicherung angewandt werden, unter Artikel 3 Nummer 1 des AI Acts hätte für die aktuarielle Praxis erhebliche, in der Sache nicht gerechtfertigte Folgen, ohne einen erkennbaren zusätzlichen Beitrag zum Verbraucherschutz, zum Wohle des Versichertenkollektivs oder zum Schutz der Grundrechte der Versicherungsnehmerinnen und Versicherungsnehmer sowie sonstiger betroffener Personen zu generieren.

Vor diesem Hintergrund fordert die DAV die Europäische Kommission auf klarzustellen, dass klassische statistische Verfahren wie GLMs und GAMs bei fachlich vorab spezifizierter, parametrisch erklärbarer und nicht selbstanpassender Ausgestaltung regelmäßig nicht als KI-Systeme im Sinne von Artikel 3 Nummer 1 des AI Acts einzuordnen sind. Die bestehende aktuarielle und aufsichtsrechtliche Governance dieser Verfahren hält sie für hinreichend, um deren spezifische Risiken zu adressieren.

Bis zu einer entsprechenden Klarstellung besteht eine gewisse Rechtsunsicherheit bei der Einordnung klassischer statistischer Verfahren im Sinne des AI Acts. Unternehmen sollten deshalb ihre

[29] Siehe Stellungnahme der Deutschen Aktuarvereinigung e. V. „DAV plädiert für eine rechtliche Klarstellung im AI Act zur Abgrenzung klassischer statistischer Verfahren in der KI-Definition“ vom 08.06.2026.

Entscheidung, GLM- und GAM-basierte Verfahren nicht als KI-Systeme einzustufen, nachvollziehbar dokumentieren. Der vorliegende Ergebnisbericht gibt hierfür fachliche Argumente, ersetzt jedoch keine Rechtsberatung.

Anhang A. Grundlagen des maschinellen Lernens

Dieser Anhang vermittelt die Grundlagen des überwachten maschinellen Lernens, stellt die im Bericht genannten Prognoseverfahren und Werkzeuge vor und geht auf Erklärbarkeit sowie die technische Umsetzung ein.

A.1. Einführung ins überwachte maschinelle Lernen

Beim überwachten maschinellen Lernen werden aus Eingabedaten datengetrieben Vorhersagemodelle erstellt. Kennzeichnend ist, dass zu jeder Beobachtung eine bekannte Zielgröße – die gewünschte Ausgabe – vorliegt, anhand derer das Modell den Zusammenhang zwischen den Eingaben und der Ausgabe erlernt – anders als beim unüberwachten Lernen, das ohne eine solche Zielgröße Strukturen in den Daten aufdeckt. Je nach Art der Zielgröße unterscheidet man zwei Grundklassen von Problemstellungen: die Regression mit stetiger Zielgröße, etwa die Vorhersage eines Schadenbetrags, und die Klassifikation mit kategorialer Zielgröße, die sich weiter in binäre Klassifikation (zwei mögliche Klassen, etwa Storno ja/nein) und Mehrklassen-Klassifikation unterteilen lässt.

Die im Folgenden genannten Methoden und Tools werden exemplarisch in den in Kapitel 5 beschriebenen Anwendungsfällen verwendet und dort teils näher beschrieben.

A.1.1. Generalisierung und Overfitting

Ein zentrales Ziel ist die Generalisierung, also die Fähigkeit eines Modells, auch auf neuen, bislang unbekannten Daten gute Vorhersagen zu treffen. Voraussetzung hierfür ist, dass diese Daten aus derselben oder einer zumindest hinreichend ähnlichen Verteilung wie die Trainingsdaten stammen. Ein gut generalisierendes Modell erfasst die relevanten Muster und Zusammenhänge in den Trainingsdaten, ohne sich an zufällige Besonderheiten der Stichprobe anzupassen. Diesem Ziel stehen zwei entgegengesetzte Phänomene gegenüber: Overfitting tritt auf, wenn ein Modell die Trainingsdaten zu gut „lernt“, also zu viele Details und zufälliges Rauschen aufnimmt; es erreicht dann auf den Trainingsdaten eine hohe Genauigkeit, verschlechtert sich aber auf neuen Daten, weil die Generalisierungsfähigkeit eingeschränkt ist. Underfitting bezeichnet den umgekehrten Fall: Ein zu einfaches Modell erfasst die zugrunde liegenden Zusammenhänge nicht ausreichend und liefert bereits auf den Trainingsdaten schwache Vorhersagen.

Es geht also darum, die datengerechte Modellkomplexität zwischen diesen beiden Extremen zu finden. Um eine passende Modellkomplexität sicherzustellen, teilt man daher den Datensatz in (mindestens) Trainingsdaten und Validierungsdaten ein. Mit zunehmender Modellkomplexität steigt die Genauigkeit bei den Trainingsdaten, kann aber bei den Validierungsdaten abnehmen, was dann ein Zeichen für Overfitting ist.

Strategien wie Regularisierung (siehe Abschnitt A.2.2) und Kreuzvalidierung können helfen, Overfitting zu begrenzen und die Generalisierung zu verbessern.

A.1.2. Datenaufbereitung und Visualisierung

Wie bei klassischen Regressionsverfahren müssen auch bei vielen Verfahren des maschinellen Lernens fehlende Werte geeignet behandelt und nominale Merkmale in eine numerisch verarbeitbare Form überführt werden. Darüber hinaus erfordern einige Algorithmen eine Skalierung oder Standardisierung der Merkmalswerte, um eine stabile und effiziente Modellschätzung zu gewährleisten.

Je nach Datensatz kann die Datenaufbereitung durch Feature Engineering wesentlich zur Verbesserung der Modellgüte beitragen. Dabei werden beispielsweise neue Merkmale aus bestehenden Variablen abgeleitet, etwa durch die Bildung von Verhältniszahlen („Risikoabbildung“) oder die Einbindung zusätzlicher Datenquellen.

Falls die Validierungsdaten für einzelne Modellentscheidungen verwendet werden, so eignen sie sich nicht mehr für die endgültige Modellbewertung. Zu diesem Zweck werden dann Testdaten

verwendet. Der Datensatz wird daher typischerweise in drei Stichproben für Training, Validierung und Test aufgeteilt, wobei der größte Teil der Daten für das Training verwendet wird.

Die Visualisierung der Daten unterstützt das Verständnis der Datenstruktur und erleichtert die Identifikation von Zusammenhängen, Anomalien, Ausreißern und Qualitätsmängeln. Darüber hinaus kann sie wertvolle Hinweise für die Auswahl geeigneter Aufbereitungs- und Modellierungsstrategien liefern. Hierfür liefert der Anwendungsfall „Analyse der Prämienstruktur eines Krankenversicherungsbestands“ in Abschnitt 5.5 ein anschauliches Beispiel.

Die Aufbereitung der Daten ist ein erfolgskritischer Punkt, da Fehler oder bereits Unsauberkeiten via „Data Leakage“ zu fälschlich hohen Prognosegüten in allen (gleich aufbereiteten) Stichproben führen können, die Modelle im Praxiseinsatz jedoch enttäuschen. Es ist in besonderem Maße darauf zu achten, dass bei der Datenaufbereitung den ML-Modellen nur Informationen zur Verfügung gestellt werden, die zum Anwendungszeitpunkt tatsächlich bekannt sind und in gleicher Art und Weise vorliegen.

A.1.3. Gütebewertung

Ebenfalls sehr wichtig ist die Wahl einer geeigneten Bewertungsmetrik für das betrachtete Modell. Bei binären Klassifikationsproblemen wird häufig der AUC-Wert (Area Under the Curve) verwendet. Er basiert auf der ROC-Kurve (Receiver Operating Characteristic), die die Leistung eines Klassifizierers darstellt, indem sie die Rate der echt positiven Ergebnisse gegen die Rate der falsch positiven Ergebnisse aufträgt. Der AUC-Wert bietet Vorteile gegenüber der Genauigkeit (Anteil der korrekten Klassifizierungen), da er robuster gegenüber Klassenungleichgewichten ist, sowohl falsch positive als auch falsch negative Fehlentscheidungen berücksichtigt und die Leistung über verschiedene Schwellenwerte hinweg betrachtet. Der AUC-Wert bietet daher eine präzisere Beurteilung der Klassifizierungsleistung. Bei Regressionsproblemen mit stetiger Zielgröße kommen dagegen Fehlermaße wie der mittlere quadratische Fehler (MSE) bzw. dessen Wurzel (RMSE) oder der mittlere absolute Fehler (MAE) zum Einsatz, die die Abweichung zwischen vorhergesagten und tatsächlichen Werten messen. Bei der Mehrklassen-Klassifikation lassen sich Kennzahlen wie die Genauigkeit sinngemäß erweitern, etwa durch Mittelung über die einzelnen Klassen.

A.2. Prognoseverfahren

Die folgenden Abschnitte stellen die im Bericht verwendeten Prognoseverfahren vor – von Entscheidungsbäumen und Gradient Boosting über die logistische Regression bis hin zu Deep-Learning-Ansätzen und deren Kombination.

A.2.1. Entscheidungsbäume und Gradient Boosting

Ein Entscheidungsbaum ist ein Algorithmus für überwachtes maschinelles Lernen, der sowohl für Regressions- als auch für Klassifikationsaufgaben eingesetzt werden kann. Er ähnelt einem Baum mit Entscheidungsknoten, die auf Merkmalen basieren, und Endknoten (Blätter), die das Ergebnis darstellen. Der Baum wird aufgebaut, indem das Merkmal ausgewählt wird, das den größten Informationsgewinn bietet, und diese Aufteilung wird rekursiv fortgesetzt, bis bestimmte Bedingungen erfüllt sind. Entscheidungsbäume sind leicht zu verstehen, aber sie können anfällig für Overfitting und Underfitting sein.

Beim „Gradient Tree Boosting“ wird zunächst ein i. d. R. recht kleiner Baum berechnet und mit einer großen Anzahl von Iterationen versucht, die Residuen sukzessive zu reduzieren. Das Ergebnis wird aufsummiert, wobei ein neues Iterationsergebnis mit einem kleinen Gewicht in Höhe der sogenannten Lernrate einfließt und das Verfahren daher langsam lernt und gewissenhaft „getunt“ werden muss. Leistungsfähige Implementierungen dieses Prinzips finden sich in den Top-Tools CatBoost, LightGBM und XGBoost, die etwa in R und Python eingesetzt werden können. Während CatBoost bereits mit Standardeinstellungen zu sehr guten Prognosen führt, müssen insbesondere bei XGBoost zumindest die Lernrate und die Anzahl der Iterationen mittels eines Hyperparameter-Tunings optimiert werden.

Gradient-Tree-Boosting-Verfahren führen seit einiger Zeit zu den genauesten Prognosen auf Basis tabellarischer Daten und sind recht einfach und robust anwendbar.

A.2.2. Logistische Regression und Regularisierung

Die logistische Regression kann als Spezialfall eines Generalisierten Linearen Modells (GLM) mit Binomial-Verteilung der Zielgröße und Logit-Link dargestellt werden und liefert die Koeffizienten für eine an die Daten angepasste, meist hochdimensionale Regressionsfunktion. Sie ist das klassische Standardverfahren für binäre Klassifikation.

Die regularisierte logistische Regression ist eine Erweiterung der logistischen Regression, bei der ein „Strafterm“ zur Kostenfunktion hinzugefügt wird. Der Strafterm wird durch Kreuzvalidierung bestimmt und verhindert, dass die Koeffizienten zu groß werden und Overfitting auftritt. Die L1-Regularisierung kann dazu führen, dass einige Koeffizienten auf Null gesetzt werden, was als eine Art Merkmalsauswahl fungiert, während die L2-Regularisierung dazu führt, dass die entsprechenden Koeffizienten nahe an Null herankommen, aber nicht genau Null erreichen.

Regularisierung spielt auch in anderen Verfahren (z. B. Neuronale Netze) eine wichtige Rolle bei der Verhinderung von Overfitting.

A.2.3. Künstliche neuronale Netze und Deep Learning

Vorwärts gerichtete künstliche neuronale Netze bilden eine weitere, zum Beispiel für Schrifterkennung (Optical Character Recognition) verbreitete Klasse von ML-Verfahren. Sie sind aus Schichten von Berechnungszellen (den „Neuronen“) aufgebaut, die aus einfließenden Inputdaten oder Ausgaben vorgeschalteter Schichten mit Hilfe einer parametrisierten Rechenregel neue Interims-Ausgabewerte ermitteln und weiterleiten, welche in der letzten Schicht zu einem numerischen oder kategoriellen Ergebnis (wie etwa „Ja“/ „Nein“) zusammengeführt werden.

Das Training eines neuronalen Netzes, also die Optimierung der Berechnungsparameter für jedes Neuron, basiert auf klassischen numerischen Methoden wie Gradient Descent und wird als „Deep Learning“ bezeichnet. Generative Sprachmodelle, die im Anhang B.1 genauer beschrieben werden, bestehen aus riesigen, aufeinander aufbauenden neuronalen Netzen.

Neuronale Netze haben sich in den inzwischen zahlreichen Anwendungsfällen und Notebooks (siehe etwa die Musterlösungen zu ADS Immersion[30]), die im aktuariellen Umfeld mit tabellarischen Daten entstanden sind, als nur mäßig erfolgreich für Analysezwecke erwiesen und werden daher in den obigen Anwendungsfällen 5.4 und 5.5 nicht weiter vertieft dargestellt.

A.2.4. Blending und Stacking

Blending und Stacking zählen zu den Ensemble-Methoden, die die Vorhersagen mehrerer Modelle kombinieren, um eine höhere Treffsicherheit und Stabilität zu erreichen, als sie ein einzelnes Modell bietet. Für ein erfolgreiches Blending oder Stacking benötigt man die Vorhersagen von mehreren, möglichst guten und unterschiedlichen Modellen. Beim Blending werden die Prognosewerte zu einem Mittelwert oder einer Mehrheitsentscheidung kombiniert.

Beim Stacking werden mehrstufige Modelle erstellt, bei denen ab der zweiten Stufe nicht mehr die Trainingsdaten, sondern die Prognosen der darunter liegenden Modellstufe verwendet und zu einem neuen, potenziell besseren Vorhersagemodell verarbeitet werden. Zur Vermeidung von Overfitting wird dabei ein komplexes Stichprobenmodell verwendet.

Umfangreiche Modell-Stacks gewinnen typischerweise Prognosewettbewerbe mit tabellarischen Daten, sind inzwischen aber auch sehr einfach über AutoML-Tools erstellbar.

[30] Die ADS Immersion ist Teil der DAV-Zusatzqualifikation Certified Actuarial Data Scientist (CADS); siehe https://aktuar.de/de/ueber-uns/qualifizierung/cads-ausbildung/

A.3. Weitere Themen

Dieser Abschnitt beleuchtet ergänzende Aspekte: automatisierte Modellierungswerkzeuge, die Erklärbarkeit von Modellen sowie Fragen der software- und hardwareseitigen Umsetzung.

A.3.1. AutoML-Tools

Diese Verfahren führen weitestgehend automatisiert ML-Prognosen durch und verwenden einige der erfolgreichsten Prognosemethoden und teilweise Stacking-Methoden. Typischerweise kann ein Laufzeitbudget angegeben werden.

Ein Beispiel für ein solches Werkzeug ist AutoGluon, das alle genannten Anforderungen erfüllt. Die resultierenden Modell-Stacks können selbst bei vergleichbarer Modellgüte und auf demselben Datensatz deutlich unterschiedlich ausfallen. Dies wird beispielsweise in der letzten Aufgabe der prämierten „CADS Immersion Best Notebooks 2025“ veranschaulicht.

A.3.2. Erklärbarkeit

Mit Ausnahme von einfachen Verfahren wie der Logistischen Regression (bzw. allgemeiner: GLM- und GAM-Verfahren) oder Entscheidungsbäumen erzeugen die oben beschriebenen Verfahren sogenannte „Black-Box-Modelle“, die nicht direkt interpretierbar sind. Daher werden die Prognosen von ML-Modellen typischerweise auf globaler und lokaler Ebene mittels modellagnostischer Surrogatmodelle interpretiert und erklärt[31]. Oft werden hierzu die ursprünglich aus der Spieltheorie stammenden Shapley-Values bzw. daran orientierte Verfahren mit dem Tool SHAP verwendet. Für Näheres zur Theorie wird auf das im Internet verfügbare Buch „Interpretable Machine Learning“ von Christoph Molnar (2025) verwiesen[32].

A.3.3. Software und Hardware

In den vergangenen Jahren sind sehr leistungsfähige Programmbibliotheken für Datenanalyse, Visualisierung und das maschinelle Lernen für die Open-Source-Skriptsprachen R und insbesondere für Python entstanden.

Die zuvor beschriebenen Bibliotheken für tabellarische Daten – CatBoost, LightGBM und XGBoost – stehen sowohl in R als auch in Python zur Verfügung und lassen sich wahlweise auf der CPU herkömmlicher Rechner oder auf einer Grafikkarte (GPU) ausführen.

PyTorch ist ein Python-Framework für Deep Learning, das sich besonders für das Training und die Ausführung großer, rechenintensiver Modelle auf Grafikkarten eignet.

Ebenfalls laufzeitkritisch kann die Berechnung von Surrogatmodellen mit SHAP in R und Python sein.

[31] Siehe Ergebnisbericht „Explainable Artificial Intelligence: Ein aktueller Überblick für Aktuarinnen und Aktuare“ des Ausschusses ADS / AI in der Fassung vom 27.05.2024.

[32] Molnar, C. (2025): *Interpretable Machine Learning – A Guide for Making Black Box Models Explainable*, 3. Auflage, ISBN 978-3-911578-03-5, abrufbar unter https://christophm.github.io/interpretable-ml-book/

Anhang B. Grundlagen der generativen künstlichen Intelligenz

Dieser Anhang vermittelt die Grundlagen der generativen künstlichen Intelligenz, erläutert die Funktionsweise großer Sprachmodelle sowie ihre Steuerung und Anbindung an eigene Daten und geht auf ihre Grenzen und Rahmenbedingungen ein.

B.1. Von Text zu Sprachmodellen

B.1.1. Tokenisierung und Embeddings

Ein Sprachmodell verarbeitet keinen Fließtext, sondern eine Folge von Einheiten, die als Token bezeichnet werden. Gebräuchlich sind Teilwortverfahren, die häufige Zeichenfolgen zu einem Token zusammenfassen und seltene Wörter in mehrere Token zerlegen. Auf diese Weise bleibt das Vokabular mit typischerweise einigen zehntausend Einträgen beherrschbar, und auch bislang unbekannte Wörter lassen sich darstellen. Jedes Token wird anschließend als Vektor in einem hochdimensionalen Raum abgebildet; diese Darstellung wird als Embedding bezeichnet. Die Vektoren werden so bestimmt, dass Token mit ähnlichem Verwendungskontext nahe beieinanderliegen, sodass sich inhaltliche Ähnlichkeit rechnerisch messen lässt. Zwei Folgerungen sind dabei wesentlich: Die Darstellung beruht ausschließlich auf Regelmäßigkeiten der Trainingsdaten, sodass dort angelegte Verzerrungen übernommen werden; und die Zerlegung in Token bildet zugleich die Abrechnungseinheit kommerzieller Anbieter.

B.1.2. Vom Zählmodell zum großen Sprachmodell

Frühe Sprachmodelle schätzten die Wahrscheinlichkeit des nächsten Wortes aus der beobachteten Häufigkeit kurzer Wortfolgen. Solche Modelle sind einfach nachvollziehbar, scheitern aber an längeren Zusammenhängen, weil die Zahl möglicher Wortfolgen mit der betrachteten Länge sehr schnell wächst. Große Sprachmodelle verfolgen dasselbe Ziel – die Vorhersage des jeweils nächsten Tokens –, verfügen jedoch über eine sehr große Zahl von Parametern und berücksichtigen einen erheblich längeren Zusammenhang. Ein Text entsteht, indem das Modell diese Vorhersage wiederholt ausführt und das jeweils ausgewählte Token wieder an die Eingabe anfügt. Daraus folgt eine für die Beurteilung zentrale Eigenschaft: Ein Sprachmodell gibt keine nachgeschlagene Antwort wieder, sondern erzeugt eine sprachlich plausible Fortsetzung.

B.1.3. Transformer und Attention

Die Leistungsfähigkeit heutiger Sprachmodelle beruht auf der im Jahr 2017 vorgestellten Transformer-Architektur. Ihr Kern ist der Attention-Mechanismus: Für jedes Token bewertet das Modell, welche anderen Token der Eingabe für dessen Bedeutung im konkreten Zusammenhang wesentlich sind, und bildet daraus eine kontextabhängige Darstellung. Da diese Bewertung für alle Positionen zugleich erfolgt, lässt sich das Training weitgehend parallelisieren; erst dies hat Modelle in der heutigen Größenordnung ermöglicht. So ist beispielsweise die Modellgröße von ca. 117 Millionen Parametern im Jahr 2018 (GPT-1) auf ca. 175 Milliarden Parameter im Jahr 2020 (GPT-3; GPT-3.5 ist Basis von ChatGPT 2022) gestiegen. Aktuelle Modelle wie Claude Mythos (2026) der Firma Anthropic umfassen nach Expertenschätzungen bis zu 10 Billionen Parameter. Die Bezeichnung GPT steht für ein generatives, vorab trainiertes Transformer-Modell. Verwandte Architekturen werden auch außerhalb der Sprachverarbeitung eingesetzt, etwa für Bilder; hierauf beruhen die in Abschnitt 5.2 beschriebenen bildbasierten Prüfschritte.

B.2. Steuerung und Anbindung an eigene Daten

B.2.1. Training, Fine-Tuning und Antwortsteuerung

Die Entwicklung eines Sprachmodells verläuft mehrstufig. Auf ein umfangreiches Vortraining folgen ein überwachtes Fine-Tuning anhand beispielhafter Aufgabenlösungen sowie Verfahren, die das Modell an menschlichen Präferenzen ausrichten. Für den Einsatz stehen mehrere Stellschrauben zur Verfügung. Die Aufgabenbeschreibung in der Eingabe – der Prompt – bestimmt maßgeblich die Qualität der Antwort. Beim Erzeugen der Ausgabe steuert insbesondere die sogenannte

Temperatur, wie stark das Modell von der jeweils wahrscheinlichsten Fortsetzung abweicht; niedrige Werte begünstigen reproduzierbare, hohe Werte vielfältigere Antworten. Eine vollständige Reproduzierbarkeit ist gleichwohl nicht zugesichert. Verfügt ein Unternehmen über hinreichend viele geeignete Beispiele, kann es ein Modell zusätzlich auf eigene Aufgaben feinabstimmen; dies prägt das Verhalten des Modells, erweitert aber nicht seinen Wissensstand.

B.2.2. Retrieval-Augmented Generation

Soll ein Sprachmodell auf unternehmenseigene oder tagesaktuelle Inhalte zugreifen, ist ein erneutes Training regelmäßig weder erforderlich noch wirtschaftlich. Stattdessen werden die betreffenden Dokumente in Abschnitte zerlegt und als Embeddings abgelegt. Zu einer Anfrage werden die inhaltlich ähnlichsten Abschnitte gesucht und der Eingabe beigefügt, bevor das Modell antwortet; dieses Vorgehen wird als Retrieval-Augmented Generation (RAG) bezeichnet. Sein wesentlicher Vorteil liegt neben der Aktualität in der Nachvollziehbarkeit, da sich die herangezogenen Fundstellen ausweisen und prüfen lassen. Die Güte hängt allerdings unmittelbar von der Zerlegung der Dokumente und von der Suche ab: Werden die einschlägigen Abschnitte nicht gefunden, antwortet das Modell gleichwohl. Als Faustregel gilt, dass Retrieval-Augmented Generation das verfügbare Wissen erweitert, während die Feinabstimmung das Verhalten prägt.

B.2.3. Strukturierte Ausgaben, Werkzeugaufrufe und Agenten

Für die Einbindung in aktuarielle Verfahren sind zwei weitere Techniken von Bedeutung. Zum einen lässt sich die Ausgabe auf ein vorgegebenes Schema festlegen, sodass das Modell keinen Fließtext, sondern eine maschinenlesbare und damit unmittelbar prüfbare Struktur liefert. Zum anderen kann das Modell festgelegte Funktionen anstoßen, deren Berechnung außerhalb des Modells deterministisch abläuft; das Modell wählt lediglich den Aufruf und die Übergabewerte aus. Beide Techniken verlagern die eigentliche Rechenarbeit aus dem probabilistischen in einen prüfbaren Teil und erzeugen zugleich einen protokollierbaren Ablauf. Das in Abschnitt 5.2 beschriebene Prüfverfahren nutzt genau dieses Muster. Werden mehrere solcher Werkzeuge zu einem eigenständig planenden System verbunden, handelt es sich um KI-Agenten im Sinne des Abschnitts 2.4.

B.3. Grenzen und Rahmenbedingungen

B.3.1. Halluzinationen, Verzerrungen und Nachvollziehbarkeit

Da ein Sprachmodell eine plausible Fortsetzung erzeugt und keine gesicherte Auskunft erteilt, kann es sachlich unzutreffende Aussagen sprachlich überzeugend formulieren. Solche Halluzinationen sind kein Randphänomen, sondern eine Eigenschaft des Verfahrens; sie lassen sich durch Quellenanbindung, strukturierte Ausgaben und deterministische Nachrechnung verringern, aber nicht ausschließen. Hinzu kommt, dass die Ausgabe von der Formulierung der Eingabe abhängt und bei wiederholter Anwendung abweichen kann. Für strukturierte, tabellarische Fragestellungen sind die in Anhang A beschriebenen Verfahren daher in aller Regel stabiler, sparsamer und besser prüfbar; die Stärke generativer Modelle liegt in der Verarbeitung von Sprache.

B.3.2. Datenschutz, Vertraulichkeit und Sicherheit

Werden Sprachmodelle als Dienst eines Dritten genutzt, verlassen die übermittelten Inhalte das Unternehmen. In der Krankenversicherung ist dies wegen der besonderen Kategorien personenbezogener Daten von erheblicher Bedeutung; die Anforderungen aus Abschnitt 3.4.1 gelten unverändert. Hinzu tritt eine verfahrensspezifische Gefährdung: Da Anweisung und verarbeiteter Inhalt in derselben Eingabe stehen, können in einem Dokument enthaltene Formulierungen das Verhalten des Modells beeinflussen. Bei der Verarbeitung fremder Unterlagen ist dies bei der Ausgestaltung der Kontrollen zu berücksichtigen. Die regulatorische Einordnung richtet sich nach Kapitel 3.

B.3.3. Kosten, Ressourcen und Einsatzentscheidung

Kommerzielle Anbieter rechnen nach verarbeiteten Token ab, sodass die laufenden Kosten mit Menge und Länge der Anfragen steigen; leistungsfähigere Modelle sind dabei regelmäßig teurer und langsamer. Für eine wirtschaftliche Beurteilung sind neben den Entgelten auch der Aufwand für Anbindung, Prüfung und Betrieb zu berücksichtigen. Die Modellwahl ist überdies selbst Gegenstand der Validierung: Abschnitt 5.2 zeigt, dass ein kleineres Modell Befunde übersehen kann, die ein leistungsfähigeres erkennt. Vor dem Einsatz ist daher zu prüfen, ob die Aufgabe überhaupt sprachlicher Natur ist und ob ein deterministisches Verfahren sie nicht einfacher, günstiger und nachvollziehbarer löst.

Anhang C. Abkürzungsverzeichnis

Im vorliegenden Ergebnisbericht werden folgende Abkürzungen verwendet:

Abkürzung	Bedeutung
ADS	Actuarial Data Science
AF	Auslösender Faktor
AGG	Allgemeines Gleichbehandlungsgesetz
AI	Artificial Intelligence (Künstliche Intelligenz)
AI Act	EU-Verordnung über Künstliche Intelligenz
ARIMA	AutoRegressive Integrated Moving Average (Zeitreihenmodell)
AUC	Area Under the Curve (Fläche unter der ROC-Kurve)
AVB	Allgemeine Versicherungsbedingungen
Bafin	Bundesanstalt für Finanzdienstleistungsaufsicht
DORA	Digital Operational Resilience Act (EU-Verordnung)
DSGVO	Datenschutz-Grundverordnung
EDV	Elektronische Datenverarbeitung
GAM	Generalisiertes Additives Modell (Generalised Additive Model)
GKV	Gesetzliche Krankenversicherung
GLM	Generalisiertes Lineares Modell (Generalised Linear Model)
GPT	Generative Pretrained Transformer
GUI	Graphical User Interface (grafische Bedienoberfläche)
ICD	International Statistical Classification of Diseases and Related Health Problems
IKT	Informations- und Kommunikationstechnologien
KI	Künstliche Intelligenz
KV	Krankenversicherung
KVAV	Krankenversicherungsaufsichtsverordnung
LLM	Large Language Model (großes Sprachmodell)
ML	Maschinelles Lernen (Machine Learning)
Morbi-RSA	Morbiditätsorientierter Risikostrukturausgleich (GKV)
NLP	Natural Language Processing (automatische Verarbeitung natürlicher Sprache)
PKV	Private Krankenversicherung
RAG	Retrieval-Augmented Generation
RfB	Rückstellung für Beitragsrückerstattung
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations (Verfahren zur Erklärbarkeit von ML-Modellen)
TB	Technische Berechnungsgrundlage(n)
VAG	Versicherungsaufsichtsgesetz

Abkürzung	Bedeutung
VAIT	Versicherungsaufsichtliche Anforderungen an die IT
VBA	Visual Basic for Applications (Programmiersprache)
VVG	Versicherungsvertragsgesetz