Ergebnisbericht (gemäß Nr. 14.1 ANBest-IF)

Konsortialführung:	aQua - Institut für angewandte Qualitätsförderung und Forschung im Gesundheitswesen GmbH
Förderkennzeichen:	01VSF20014
Akronym:	KI-THRUST
Projekttitel:	Potenziale KI-gestützter Vorhersageverfahren auf Basis von Routinedaten
Autorinnen und Autoren:	Thorsten Pollmann, Matthias Kretzler, David Ramcke, Miriam C. Maurer, Zully Ritter, Jacqueline M. Metsch, Lisa Weller, Anne-Christin Hauschild, Thomas G. Grobe
Förderzeitraum:	1. Juli 2021 – 31. Dezember 2024
Ansprechpartner:	Dr. Thomas Grobe E-Mail: thomas.grobe@aqua-institut.de

Das dieser Veröffentlichung zugrundliegende Projekt KI-THRUST wurde mit Mitteln des Innovationsausschusses beim Gemeinsamen Bundesausschuss unter dem Förderkennzeichen 01VSF20014 gefördert.

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Zusammenfassung

Hintergrund: Im Projekt KI-THRUST (01VSF20014) wird untersucht, wie gut sich strukturierte Routinedaten der gesetzlichen Krankenversicherung (GKV) mit Verfahren der Künstlichen Intelligenz (KI) aus dem Bereich des Maschinellen Lernens (ML) analysieren lassen. Dazu werden im Projekt unterschiedliche KI-gestützte Prädiktionsmodelle entwickelt und hinsichtlich Ihrer Eignung zur Vorhersage von zwei exemplarischen Outcomes (Ungeplante Wiederaufnahme, Mortalität) mit klassischen Regressionsanalysen verglichen. Zur Beurteilung werden neben der Modellgüte auch Aspekte zur Erklärbarkeit, Übertragbarkeit und Implementierbarkeit der Modelle berücksichtigt. Zudem ist es Ziel des Projektes, ein anwendungsorientiertes Weißbuch zu erstellen, das einen Einblick in GKV-Routinedaten gibt und Ansätze zum Einsatz von KI-Techniken vermittelt.

Methodik: Grundlage bilden Routinedaten der vier projektbeteiligten Krankenkassen (BAHN-BKK, Novitas BKK, Pronova BKK und SBK) zu ca. 1,4 Mio. Versicherten mit mindestens einer Entlassung aus dem Krankenhaus in den Jahren 2015 bis 2020. Bereitgestellt wurden Versichertenstammdaten, ambulante und stationäre Daten, Arzneimittel- und Hilfsmitteldaten sowie Pflegedaten. Nach der Datenaufbereitung und einem Train-Test-Split (80:20) wurden mit Daten aus dem Jahr 2018 unterschiedliche ML-Modelle (Random Forest, Neuronale Netze, Adaptive Boosting) sowie Regressionsmodelle trainiert. Anschließend wurde die Performance der Modelle anhand etablierter Metriken (AUC-ROC, AUC-PR) auf Basis der Testdaten 2018 verglichen. In weiteren Post-hoc-Analysen wurden Erklärbarkeitsansätze, die Übertragbarkeit auf „künftige“ Jahre (2019, 2020) sowie die Neigung zu Fehlklassifikationen in Subgruppen getestet.

Ergebnisse: Die Modellevaluation ergab, dass auf Basis identischer Daten das ML-Verfahren AdaBoost (AB) und die logistische Regression (LR) die höchste Modellgüte (gem. AUC-ROC) bei der Vorhersage der Mortalität (AB=0,889 LR=0,888) und ungeplanten Wiederaufnahme (AB=0,694 LR=0,693) erzielen. Dabei ließ sich die Mortalität besser vorhersagen als die Wiederaufnahme, wobei niedrige AUC-PR-Werte auf eine eingeschränkte Nutzbarkeit für die Individualprognostik hinweisen. Die Testung der trainierten Modelle mit Daten aus 2019 und 2020 führte trotz COVID-19-Pandemie nur zu einer leicht reduzierten Vorhersagegenauigkeit. Zudem weisen Subgruppenanalysen darauf hin, dass alle Modelle in kleineren Hochrisikogruppen (z.B. 80+ Jährige) geringfügig schlechter performen. Die im Projekt getesteten ML-Erklärungsansätze (z.B. LIME, Shapley Value) eignen sich aktuell noch nicht, um hinreichend erklären zu können, warum eine Person eine bestimmte Vorhersage erhält.

Diskussion: Beim Vergleich der ML- und Regressionsverfahren ist festzuhalten, dass mit identischen Daten eine vergleichbare Performance erreicht wird. Hier besteht weiterer Forschungsbedarf, inwiefern sich Methoden zur Datenverdichtung (z.B. Transformer) eignen, um der Komplexität von GKV-Routinedaten gerecht zu werden und das Potenzial von ML-Verfahren voll auszuschöpfen. Grundsätzlich erwiesen sich die Routinedaten für den Einsatz mit ML-Verfahren als geeignet, wobei ihre Vorteile (Datenqualität, Umfang) und Nachteile (Datenverzug, rechtliche Auflagen) in Hinblick auf das Modelltraining und die spätere Anwendung in der Praxis abzuwägen sind. Hierzu wurden praktische Umsetzungs- und Handlungsempfehlungen in einem Weißbuch beschrieben und veröffentlicht.

Schlagworte: Künstliche Intelligenz, Maschinelles Lernen, Prognosemodell, GKV-Routinedaten, Weißbuch

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Inhaltsverzeichnis

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

I Abkürzungsverzeichnis

AB	Adaptive Boosting, kurz AdaBoost
AI	Artifical Intelligence (dt. Künstliche Intelligenz)
API	Application Programming Interface (dt. Programmierschnittstelle)
BAS	Bundesamt für Soziale Sicherung
BSNR	Betriebsstättennummer
EM	Entlassmanagement
ePA	elektronische Patientenakte
G-BA	Gemeinsamer Bundesausschuss
GDNG	Gesundheitsdatennutzungsgesetz
GKV	Gesetzliche Krankenversicherung
ICD	International Classification of Diseases
IK	Institutionskennzeichen
KIS	Krankenhausinformationssystem
KNN	Künstliches Neuronales Netz
LR	Logistische Regression
ML	Machine Learning
MLP	Multi-Layer-Perzeptron
PPV	Positive Predictive Value (dt. Positive Vorhersagekraft)
PR	Precision-Recall
RF	Random Forest
ROC	Receiver-Operating-Characteristic
SGB	Sozialgesetzbuch
SOP	Standard Operating Procedure
VERSID	Versichertennummer
XAI	Explainable AI (dt. Erklärbare KI)

II Abbildungsverzeichnis

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

III Tabellenverzeichnis

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

1 Projektziele

1.1 Hintergrund

Die Digitalisierung und der zunehmende Einsatz der Künstlichen Intelligenz (KI) verändern rasant das Gesundheitswesen in Deutschland. So ermöglichen vielfältige Anwendungsmöglichkeiten aus dem Bereich der KI nicht nur die Verbesserung der Diagnostik und Therapie von Patientinnen und Patienten, sondern auch die Optimierung von Strukturen und Prozessen in der Gesundheitsversorgung, beispielsweise durch die Unterstützung der Leistungserbringenden bei ihrer alltäglichen Arbeit. Ein konkreter Anwendungszweck ist hierbei die KI-gestützte Vorhersage von gesundheitsbezogenen Ereignissen (z. B. das Eintreten einer Erkrankung) oder spezifischen Versorgungsbedarfen (z. B. das frühzeitige Erkennen eines Pflegebedarfs). Für den Einsatz von KI zu prognostischen Zwecken werden in der Regel umfangreiche und zum jetzigen Zeitpunkt noch vornehmlich strukturierte Gesundheitsdaten benötigt.

Ein nicht unwesentlicher Teil gesundheitsbezogener Daten wird in Deutschland primär zu Abrechnungszwecken und zur Prüfung und Abwicklung von Versicherungsleistungen der gesetzlichen Krankenversicherung (GKV) erfasst, über die in Deutschland die gesundheitliche Versorgung von mehr als 85 Prozent der Bevölkerung abgesichert ist. Um die Abrechnung einer Vielzahl an Leistungserbringern (wie Arztpraxen, Krankenhäusern und Apotheken) mit einer hohen zweistelligen Zahl an gesetzlichen Krankenkassen zu erleichtern und bestimmte Berichtspflichten der GKV-Krankenkassen zu ermöglichen, sind Formate und Umfänge der Datenübermittlung in vielen Bereichen bundesweit einheitlich geregelt. So verfügen alle GKV-Krankenkassen über strukturierte und vergleichbare Informationen zu den jeweils bei ihnen versicherten Personen. Diese Informationen lassen sich zusammenfassend als „GKV-Routinedaten“ bezeichnen.

In diesem Kontext verbindet das Innovationsfondsprojekt KI-THRUST die beiden Bereiche „Künstliche Intelligenz“ und „GKV-Routinedaten“ und untersucht anhand eines versorgungsrelevanten Anwendungsbeispiels die Potenziale von KI-gestützten Vorhersageverfahren, die auf der Basis von Routinedaten bei Betriebskrankenkassen entwickelt und evaluiert werden. Dabei baut das Projekt KI-THRUST auf zwei abgeschlossene Innovationsfondsprojekte auf, an denen das aQua-Institut und der BKK Dachverband bereits beteiligt gewesen sind. So wurden im Projekt EMSE (Förderkennzeichen 01VSF16041) routinedatenbasierte Prognosemodelle für das Entlassmanagement (EM) im Krankenhaus entwickelt, um bereits zu Beginn eines stationären Aufenthaltes verschiedene Nachsorgebedarfe vorhersagen und frühzeitig Maßnahmen im EM einleiten zu können (Broge et al., 2020). Im Rahmen des Projektes USER (Förderkennzeichen 01NVF18010) wurden die Prognosemodelle in Form eines Entscheidungsunterstützungssystems an verschiedenen Krankenhausstandorten in Nordrhein-Westfalen erprobt und evaluiert (Broge et al., 2024). Bei den eingesetzten Prognosemodellen handelte es sich um logistische Regressionsmodelle, die als etablierte und robuste Modellierungsverfahren der klassischen Statistik gelten und seit Jahrzenten in der Wissenschaft und Forschung verwendet werden. An dieser Stelle setzt das Projekt KI-THRUST an und untersucht, inwiefern sich modernere Verfahren der Künstlichen Intelligenz, insbesondere aus dem Bereich des Maschinellen Lernens (engl. Machine Learning, ML), grundsätzlich auf GKV-Routinedaten anwenden lassen und gegebenenfalls in der Lage sind, präzisere Vorhersagen zu erzielen als traditionelle Regressionsmodelle.

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

1.2 Ziele und Fragestellungen

Vor dem geschilderten Hintergrund besteht das übergeordnete Ziel von KI-THRUST darin, geeignete KI-Verfahren zur Prädiktion poststationärer Ereignisse auszuwählen, diese auf der Grundlage von GKV-Routinedaten zu entwickeln und anschließend hinsichtlich ihrer Möglichkeiten und Limitationen im Vergleich zu klassischen Vorhersagemodellen zu untersuchen (Projektziel 1). Hierbei sollen zudem die Eignung von GKV-Routinedaten für KI-gestützte Analysetechniken geprüft und mögliche Besonderheiten und Erfordernisse bei der Datenaufbereitung herausgearbeitet werden. Beim Vergleich der klassischen und KI-gestützten Verfahren (Projektziel 2) wird die Vorhersage- bzw. Modellgüte als primäres Vergleichskriterium herangezogen, da die Genauigkeit der Vorhersagen eine wesentliche Voraussetzung für die Praktikabilität der Prognosemodelle in verschiedenen Anwendungsszenarien der Regelversorgung darstellt. Darüber hinaus fließen weitere Aspekte, wie die Erklärbarkeit und Übertragbarkeit der Modelle oder auch Unterschiede bei den technischen Voraussetzungen, in die Bewertung mit ein.

Neben der analytischen Zielsetzung verfolgt das Projekt zudem das Ziel, die aus den Analysen gewonnenen Erkenntnisse für ein möglichst breites Publikum in verständlicher Form aufzubereiten und zugänglich zu machen. Hierbei sollen vor allem interessierten Akteuren aus den Bereichen Künstliche Intelligenz und GKV-Routinedaten praxisnahe Beispiele und Handlungsempfehlungen an die Hand gegeben werden. Dies erfolgt in Form eines Weißbuches, das während des Projektes – begleitend zu den analytischen Arbeiten – verfasst und am Projektende veröffentlicht wird (Projektziel 3). Sämtliche Projektziele und Unterziele sowie die dazugehörigen Forschungsfragen können der Tabelle 1 entnommen werden.

Tabelle 1: Projektziele und Fragestellungen

Nr.	Ziel / Unterziel	Forschungsfrage(n)
1.	**Entwicklung KI-gestützter Prädiktionsmodelle auf Basis von Routinedaten**	Allgemein: Welche Anforderungen bestehen bei der Nutzung von KI-Verfahren mit Routinedaten?
1.1.	KI-taugliche Aufbereitung der Routinedaten	Wie müssen Krankenkassendaten aufbereitet werden, damit sie für KI-Verfahren genutzt werden können?
1.2.	Identifizierung, Abgrenzung und Auswahl von geeigneten erklärbaren KI-Verfahren	Welche erklärbaren KI-Verfahren eignen sich für die Analyse von Routinedaten bei Krankenkassen? Welche Methoden der Erklärbarkeit können angewendet werden?
1.3.	Durchführung des Modelltrainings und Testung	Welche Merkmale erklären maßgeblich das Eintreten der Outcomes? Inwiefern verändert sich die Prädiktion, wenn das Timing eines Ereignisses in den Modellen mitberücksichtigt wird?
1.4.	Darstellung der Ergebnisse	Wie lassen sich die Ergebnisse bzw. die Vorhersagen transparent sowohl anwender- als auch nutzenorientiert darstellen?
2.	**Vergleich von KI- und regressionsgestützten Prädiktionsmodellen**	Welche Unterschiede und Gemeinsamkeiten bestehen zwischen routinedatenbasierten KI- und Regressionsverfahren?
2.1.	Spezifizierung der Anforderungen an konventionelle und KI-basierte Modelle	Bedarf es für unterschiedliche Modellansätze unterschiedlicher Formen der Datenaufbereitung und wenn ja, welche? Unterscheiden sich die Hard- und

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

		Softwarevoraussetzungen je nach Ansatz? Was sind Mindestanforderungen?
2.2.	Vergleich der KI- und Regressionsmodelle hinsichtlich der Qualität und Zuverlässigkeit der Vorhersagen	Unterscheiden sich die jeweiligen Modelle in der Präzision der Vorhersage von Ereignissen? Welche Rolle spielen fehlerhafte Vorhersageergebnisse – auch in Hinblick auf bestimmte Datenkonstellationen/Subgruppen?
2.3.	Bewertung der Stärken und Schwächen der KI- und Regressionsmodelle beim Einsatz in der Gesundheitsversorgung	Wann eignet sich welches Verfahren? Welche erklärbaren KI-Verfahren sind für die Analyse von Routinedaten geeignet? Gibt es Limitationen in Bezug auf den Einsatz und die Erklärbarkeit?
3.	**Ergebnissynthese in Form eines Weißbuchs**	Welche grundsätzlichen und praxisorientierten Handlungsempfehlungen sind zu erörtern, damit routinedatenbasierte KI-Verfahren an geeigneten Stellen von einem möglichst breiten Anwenderkreis eingesetzt werden können?

2 Projektdurchführung

2.1 Projektbeteiligte

Tabelle 2: Projektkonsortium

Einrichtung	Rolle im Projekt	Projektleitung	Verantwortungsbereich
aQua-Institut für angewandte Qualitätsförderung und Forschung im Gesundheitswesen GmbH (aQua-Institut)	Konsortialführung	Dr. Thomas G. Grobe	Konsortialführung, Projektsteuerung, Bereitstellung einer sicheren und KI-tauglichen Analyseumgebung, Aufbereitung und Analyse von Routinedaten, Berichtserstellung (federführend) und Publikation
Institut für Med. Informatik, Universitätsmedizin Göttingen (UMG)	Konsortialpartner	Prof. Dr. Dagmar Krefting (bis einschl. 2022), Jun.-Prof. Dr. Anne-Christin Hauschild (ab 2023)	Datenanalyse mittels KI-Verfahren, Berichtserstellung und Publikation
BKK Dachverband e.V.	Konsortialpartner	Matthias Kretzler	Koordinierungsfunktion (ggü. Betriebskrankenkassen), Öffentlichkeitsarbeit, Berichtserstellung und Publikation

Die Betriebskrankenkassen BAHN-BKK, Novitas BKK, Pronova BKK und Siemens-Betriebskrankenkasse waren an dem Projekt als Kooperationspartner beteiligt und haben ihre Routinedaten für die im Projekt vorgesehenen Datenanalysen zur Verfügung gestellt. Zudem unterstützten sie die Konsortialpartner im Austausch zu verschiedenen fachlichen Fragestellungen. Die Datenbereitstellung erfolgte über den gemeinsamen IT-Dienstleister der Betriebskrankenkassen, der BITMARCK Service GmbH. Darüber hinaus wurde das Projektkonsortium von einem wissenschaftlichen Beirat, paritätisch zusammengesetzt aus

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

Expertinnen und Experten aus den Bereichen Künstliche Intelligenz sowie Gesundheitsversorgung und -daten, begleitet und unterstützt.

2.2 Beschreibung/ Darstellung des Projekts

Um die unter Kapitel 1.2 dargelegten Projektziele zu erreichen, besteht das Projekt zunächst aus einem empirisch-analytischen Teil, im Rahmen dessen verschiedene Vorhersageverfahren des Maschinellen Lernens und konventionelle Regressionsanalysen auf Grundlage der Routinedaten der projektteilnehmenden Betriebskrankenkassen entwickelt und miteinander verglichen werden. Der Modellvergleich erfolgt primär anhand von geeigneten Evaluationsmetriken, die die Vorhersagegüte quantifizieren. Darüber hinaus werden Aspekte zur Erklärbarkeit der Modelle, zur Übertragbarkeit der Modelle auf andere Datenjahre, zum Ausmaß und zur Charakterisierung fehlklassifizierter Fälle sowie Implikationen für eine spätere Implementierung der Verfahren (z. B. technische Umsetzbarkeit, Rechenleistung) beleuchtet. In Anlehnung an die Vorläuferprojekte EMSE und USER zum Thema Entlassmanagement (siehe Kapitel 1.1), werden auch im Projekt KI-THRUST zwei poststationäre Events als exemplarische Modelloutcomes – im ML-Bereich auch „Targets“ genannt – vorhergesagt: (1.) Mortalität und (2.) Ungeplante Wiederaufnahme. Beide Outcomes zeichnen sich dadurch aus, dass sie auch abseits des Entlassmanagements eine hohe Versorgungsrelevanz aufweisen und daher häufig als Kennzahlen in der Versorgungsforschung verwendet werden. Darüber hinaus haben die Projekte EMSE und USER gezeigt, dass die beiden Outcomes unterschiedliche Eigenschaften (z. B. Prävalenz) aufweisen, die sich auf die Prognostizierbarkeit auswirken und andere Anforderungen an die Modellierung stellen. Das methodische Vorgehen im analytischen Teil des Projektes wird in Kapitel 3 detaillierter beschrieben.

Ein weiterer Bestandteil ist die projektbegleitende Erstellung eines Weißbuchs zu den Potenzialen KI-gestützter Vorhersageverfahren auf Basis von GKV-Routinedaten. In das Weißbuch fließen die oben erörterten Regressions- und ML-Analysen als praxisnahes Anwendungsbeispiel ein. Dabei wird das methodische Vorgehen Schritt für Schritt erläutert und anhand von Code-Beispielen, verwendeten Software-Packages und weiterführenden Quellen für die interessierte Leserschaft verständlich und für deren eigene Analysevorhaben nutzbar gemacht. Darüber hinaus soll das Weißbuch einen Einstieg in die Welt der GKV-Routinedaten ermöglichen. Hierzu werden nicht nur der Umfang und die Struktur entlang der einzelnen GKV-Datenbestände (z. B. Leistungsdaten von Krankenhäusern oder aus dem vertragsärztlichen Bereich) und die verwendeten Klassifikationssysteme beschrieben, sondern auch zusätzliche Aspekte thematisiert, wie die datenschutzrechtlichen Vorgaben oder mögliche Fallstricke bei der Nutzung der Daten im Forschungskontext. Am Ende des Weißbuchs werden aus den Ergebnissen Handlungsempfehlungen abgeleitet und mögliche Anwendungsgebiete für routinedatengestützte Vorhersageverfahren skizziert. Das Weißbuch soll als Online-Publikation veröffentlicht werden und frei zugänglich sein.

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

2.3 Beschreibung Ablauf des Projekts

2.3.1 Vorbereitungsphase

Zu Beginn des Projektes wurden die erforderlichen vertraglichen Vereinbarungen, darunter ein Konsortialvertrag zwischen den Konsortialpartnern sowie Kooperations- und Datenschutzverträge mit den projektbeteiligten Betriebskrankenkassen, geschlossen und ein Datenschutzkonzept erstellt. Zudem wurde ein wissenschaftlicher Beirat konstituiert, der aus Expertinnen und Experten aus den Bereichen Künstliche Intelligenz und Gesundheitsversorgung (mit Schwerpunkt Gesundheitsdaten) bestand.

Des Weiteren wurde eine Datensatzbeschreibung angefertigt und mit der BITMARCK Service GmbH hinsichtlich der Umsetzbarkeit konsentiert. In der Beschreibung werden neben dem Datenumfang auch die Ein- und Ausschlusskriterien, Vorgaben zur Pseudonymisierung, das Lieferdatum sowie der Übertragungsweg spezifiziert (siehe Anlage 3).

Am 10. Januar 2022 wurde ein Ethikantrag bei der zuständigen Ethikkommission der Ärztekammer Niedersachsen gestellt, über den die Kommission mit einem positiven Votum vom 7. Februar 2022 entschied. Darüber hinaus wurde am 17. Februar 2022 ein Antrag zur Übermittlung von Sozialdaten gemäß § 75 SGB X beim Bundesamt für Soziale Sicherung (BAS) gestellt, um die Routinedaten der Betriebskrankenkassen zu Forschungszwecken übermitteln und nutzen zu dürfen. Die Genehmigung wurde am 2. Juni 2022 durch das BAS erteilt.

Um den hohen technischen Anforderungen der rechenintensiven KI-Analysetechniken gerecht zu werden und eine Integration aktueller Technologien zu ermöglichen, wurde für das Projekt eine leistungsfähige IT-Infrastruktur benötigt. Hierzu wurde in enger Abstimmung zwischen der IT-Abteilung des aQua-Instituts und den für die Durchführung der KI-Analysen vorgesehenen Mitarbeitenden der UMG ein entsprechendes Konzept entwickelt (siehe Anlage 2). Anschließend erfolgten durch das aQua-Institut die Anschaffung (aus Mitteln des aQua-Instituts), der Aufbau und die Testung der KI-tauglichen Hardware sowie die Installation und Konfiguration der benötigten Software (inkl. Einrichtung eines gesicherten Fernzugriffs für die Projektmitarbeitenden der UMG).

2.3.2 Routinedatenanalyse

Die Projektphase bis zur Datenübermittlung wurde von den Projektbeteiligten genutzt, um den Analyseplan auszuarbeiten. In diesem Zuge erfolgte ein interdisziplinärer Wissenstransfer zwischen den Routinedatenexpertinnen und -experten des aQua-Instituts und den KI-Expertinnen und -experten der UMG, vor allem hinsichtlich des Umfangs und den Besonderheiten von GKV-Routinedaten sowie den spezifischen Datenanforderungen zur Vorbereitung der KI-Analysen. Ziel hierbei war nicht nur ein fachlicher Austausch zur Entwicklung des Analyseplans, sondern auch das Schaffen eines gemeinsamen Begriffs- und Methodenverständnisses. Darüber hinaus wurden SAS- und Python-Skripte zur Datenaufbereitung und Datenanalyse vorbereitet. Bezüglich der Datenaufbereitung und Regressionsanalysen konnte zum Teil auf Erfahrungen und Vorarbeiten aus den Vorläuferprojekten EMSE und USER (siehe Kapitel 1.1) zurückgegriffen werden. Die Vorarbeiten umfassten im Wesentlichen die Auswahl relevanter Prädiktoren (v. a. zur Vorhersage der Mortalität und der ungeplanten Wiederaufnahme), welche im Zuge eines literaturbasierten und empirischen Prozesses festgelegt worden waren, sowie einzelne SPSS-

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

und SAS-Skripte zur Datenaufbereitung und Modellierung. Die in den Vorläuferprojekten erbrachten Vorarbeiten sind in den jeweiligen Ergebnisberichten beschrieben (Broge et al., 2020; Broge et al., 2024).

Die pseudonymisierten Routinedaten wurden gemäß der Datensatzbeschreibung (siehe Anlage 3) im August 2022 von der BITMARCK Service GmbH über einen gesicherten Übertragungsweg an das aQua-Institut übermittelt. Nach einer Datenprüfung erfolgten die Datenaufbereitung und anschließende Entwicklung und Evaluation der Regressions- und KI-Analysen entlang des Analyseplans, der im Kapitel 3.2 vorgestellt wird. Sämtliche Schritte zur Datenaufbereitung der GKV-Routinedaten und die Berechnung der Regressionsmodelle wurden auf speziellen SAS-Servern des aQua-Instituts vorgenommen. Alle Schritte zur ML-spezifischen Vorverarbeitung der Daten sowie die Umsetzung der ML-Verfahren erfolgten in der oben beschriebenen Umgebung der KI-Infrastruktur durch Mitarbeitende der UMG. Trotz der Vorgaben des Analyseplans war eine gewisse Flexibilität in der Umsetzung erforderlich. Vor allem beim Trainieren der ML-Vorhersagemodelle wurden aufgrund neuer Datenkonstellationen oder Zwischenergebnisse situativ Anpassungen vorgenommen. Beispielsweise führten die für die Modellierung vorgesehenen Outcomes zu „unbalancierten Daten“ (freie Übersetzung von engl. imbalanced data), die die Integration spezieller Methoden erforderlich machten (siehe Kapitel 3.4).

2.3.3 Weißbucherstellung

Die Erstellung des Weißbuchs erfolgte projektbegleitend. Zu Beginn wurde das grundlegende Konzept des Weißbuchs und ein erster Gliederungsentwurf im Projektkonsortium erarbeitet und mit dem wissenschaftlichen Beirat abgestimmt. Anschließend wurden die Aufgaben bzw. die zu schreibenden Kapitel innerhalb des Konsortiums verteilt. Da im ersten Projektjahr noch keine Daten vorlagen und folglich keine Analysen durchgeführt werden konnten, wurde diese Zeit genutzt, um die grundlegenden Kapitel zur Einführung in die Themen GKV-Routinedaten und KI-Analysetechniken zu verfassen (siehe Anlage 1, Kapitel 1 bis 3). Die nachfolgenden Kapitel, die die praktische Umsetzung am Projektbeispiel zum Gegenstand haben (siehe Anlage 1, Kapitel 4 bis 8), wurden begleitend zur Datenaufbereitung sowie der Modellentwicklung und -evaluation verfasst. Mit der Erstellung der beiden letzten Kapitel zur Implementierung von routinedatengestützten KI-Verfahren und der Zusammenfassung der Projekterkenntnisse (siehe Anlage 1, Kapitel 9 und 10) wurde die Schreibphase abgeschlossen. Sämtliche Kapitel wurden einer internen und externen Revision, primär durch Mitglieder des wissenschaftlichen Beirats, unterzogen. In der Finalisierungsphase wurden ein internes Lektorat und die Überarbeitung des Layouts durchgeführt. Das fertige Weißbuch wurde als Online-Publikation vom aQua-Institut veröffentlicht (siehe Anlage 1).

2.3.4 Änderungen im Projektverlauf

Erste Herausforderungen ergaben sich zu Projektbeginn bei der Beschaffung der notwendigen Hardware zum Aufbau der KI-tauglichen Infrastruktur (siehe Anlage 2). Dies kollidierte zeitlich mit der sogenannten Chipkrise 2021. So führten die während der COVID-19-Pandemie verhängten Lockdowns in wichtigen Produktionsstätten, insbesondere in Asien, zu Produktionsausfällen und Störungen der Lieferketten. Dies hatte zur Folge, dass vor allem die für das Projekt benötigten Grafikkarten sehr lange Lieferzeiten aufwiesen und mit hohen Anschaffungskosten verbunden waren. Die Verzögerungen hatten jedoch keine Auswirkungen

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

auf den Projektverlauf, da die Hardware erst mit Beginn der Datenanalysen im zweiten Projektjahr benötigt wurde.

Im Vergleich dazu ging das unplanmäßig lange Genehmigungsverfahren beim BAS mit Verzögerungen bei der Datenbereitstellung und den daran anknüpfenden Projektarbeiten einher. Im Rahmen des mehrmonatigen Verfahrens wurde geprüft, inwiefern nicht nur der vorgesehene Gesamtumfang an Daten, der bei Forschungsvorhaben i. d. R. auf den minimal notwendigen Daten- und Merkmalsumfang beschränkt ist (z. B. auf bestimmte Versicherte, Leistungsbereiche, Diagnosen), sondern auch einzelne Merkmale ohne weitere Vergröberung (z. B. Sterbedatum anstatt Sterbemonat) datenschutzkonform bereitgestellt werden konnten. Hier gilt es nicht nur den aktuellen rechtlichen Vorgaben zur Nutzung der Routinedaten in der Forschung, wie dem Interessensschutz der Versicherten und dem Datensparsamkeitsgebot, sondern auch den Anforderungen von KI-Vorhaben, für deren Umsetzung möglichst umfangreiche Daten (Stichwort „Big Data“) benötigt werden, Rechnung zu tragen.

Zu Beginn des Jahres 2023 erfolgte ein Wechsel der projektverantwortlichen Person am Institut für Medizinische Informatik der UMG. So fiel das Projekt fortan in den Zuständigkeitsbereich der damaligen Arbeitsgruppe von Jun.-Prof. Dr. Anne-Christin Hauschild, in dessen Folge neue Mitarbeitende im Projekt eingesetzt wurden. Da ein Teil der Projektstellen neu besetzt werden musste und Bewerberinnen und Bewerber mit den erforderlichen KI-Kenntnissen in Anbetracht des anhaltenden KI-Hypes schwer zu finden waren, kam es phasenweise zu Verzögerungen im Projektablauf.

Aufgrund der verschiedenen, oben geschilderten Verzögerungen im Projekt wurde am 15. Januar 2024 ein Antrag auf kostenneutrale Laufzeitverlängerung um sechs Monate beim Projektträger gestellt. Mit dessen Genehmigung vom 1. März 2024 wurde der Förderzeitraum bis zum 31. Dezember 2024 verlängert.

2.4 Rechtsgrundlage

Die Förderung des Projektes KI-THRUST (Förderkennzeichen 01VSF20014) durch den Gemeinsamen Bundesausschuss erfolgt auf der gesetzlichen Grundlage gem. § 92a Abs. 2 SGB V. Grundlage hierfür sind der Förderantrag vom 20.03.2020 zur Förderbekanntmachung zur Förderung von Versorgungsforschung vom 12.12.2019 (Themenfeld 4: Perspektiven und Potenziale des Einsatzes Künstlicher Intelligenz in der Versorgung), der Förderbescheid vom 02.12.2020, die Änderungsbescheide vom 31.05.2021 und 01.03.2024 sowie die Nebenbestimmungen des Innovationsausschusses beim Gemeinsamen Bundesausschuss (G-BA) für Förderungen aus dem Innovationsfonds (ANBest-IF).

Die Verwendung von Routinedaten der gesetzlichen Krankenversicherung und Pflegeversicherung zu Forschungszwecken erfolgt mit der Genehmigung vom 03.06.2022 durch das Bundesamt für Soziale Sicherung (BAS) auf der gesetzlichen Grundlage nach § 75 SGB X (siehe auch Kapitel 2.3.1).

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

3 Methodik

3.1 Datengrundlage

Die Datengrundlage für die Entwicklung und vergleichende Testung der Regressions- und ML-Verfahren bilden pseudonymisierte Routinedaten zu ca. 1,4 Millionen Versicherten der projektteilnehmenden Betriebskrankenkassen (BAHN-BKK, Novitas BKK, Pronova BKK und Siemens-Betriebskrankenkasse) aus den Jahren 2015 bis einschließlich 2020. Der zentrale Datenabzug der BITMARCK Service GmbH enthielt Leistungsdaten zu allen Versicherten, die in dem oben definierten Beobachtungszeitraum mindestens eine Entlassung aus einer stationären Krankenhausbehandlung aufwiesen. Darüber hinaus gab es für den Datenabzug keine weiteren Ein- oder Ausschlusskriterien. Vor der Datenübermittlung wurde die Kassenzugehörigkeit verblindet und sämtliche datenschutzsensiblen Identifikatoren, darunter die Versichertennummer (VERSID), die Betriebsstättennummer (BSNR) und das Institutionskennzeichen (IK), kassenübergreifend pseudonymisiert.

Die für das Projekt bereitgestellten Routinedaten umfassten folgende GKV-Datenbestände (gemäß SGB-Kontext):

- Versichertenstammdaten (§ 284 SGB V)
- Stationäre Leistungen und Diagnosen (§ 301 SGB V)
- Ambulant-ärztliche Leistungen und Diagnosen (§ 295 SGB V)
- Arzneimittel (§ 300 SGB V)
- Heil- und Hilfsmittel (§ 302 SGB V)
- Pflegleistungen und Pflegegrade (SGB XI)

Detaillierte Angaben zum Tabellen- und Merkmalsumfang können der Datensatzbeschreibung (siehe Anlage 3) entnommen werden.

3.2 Analyseplan

Das methodische Vorgehen zur Durchführung der Regressions- und ML-Analysen ist in Abbildung 1 schematisch dargestellt und lässt sich in zwei wesentliche Schritte unterteilen.

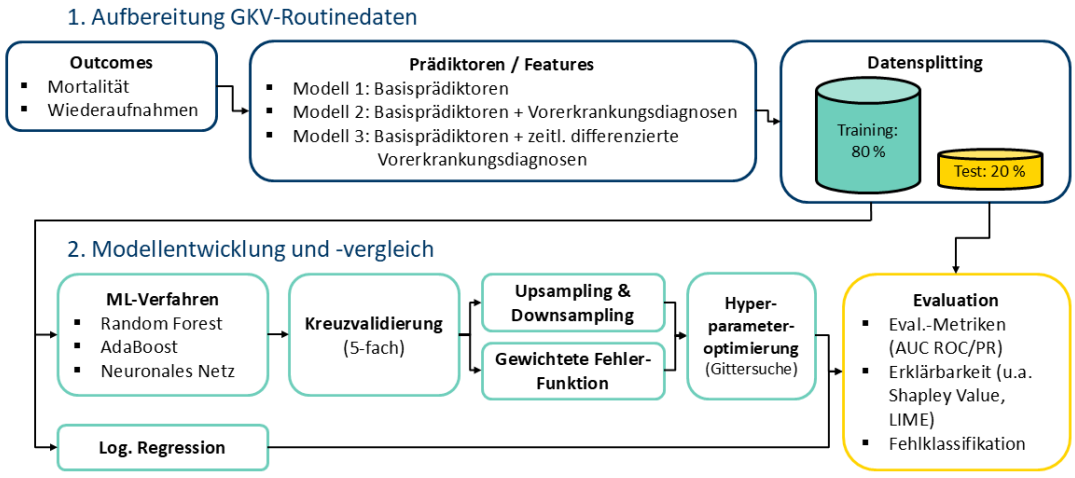

Abbildung 1: Methodisches Vorgehen bei der Modellentwicklung und -evaluation

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Der **erste Schritt (Aufbereitung GKV-Routinedaten)** umfasst alle erforderlichen Maßnahmen, um die für das Projekt bereitgestellten Rohdaten so vorzubereiten, dass sie am Ende in Form sogenannter Analysedatensätze für die nachfolgenden Modellierungen genutzt werden können. Hierzu werden nach einer grundlegenden Datenaufbereitung (siehe Kapitel 3.3.1) zunächst die beiden Outcomes Mortalität und Ungeplante Wiederaufnahme operationalisiert und berechnet (siehe Kapitel 3.3.2). Anschließend folgen die Auswahl und Berechnung der Prädiktoren (im ML-Bereich auch „Features“ genannt), denen auf der Grundlage empirischer oder theoriegeleiteter Erkenntnisse eine potenzielle Erklärungskraft für die Vorhersage der beiden Outcomes zugeschrieben wird (siehe Kapitel 3.3.3). Aufgrund der Vielzahl an Prädiktoren wurden diese im Projekt KI-THRUST in Blöcke unterteilt, die wiederum in unterschiedlichen Konstellationen insgesamt drei Datenmodelle (M1 bis M3) bilden. Die Datenaufbereitung wird abgeschlossen mit einer randomisierten Zuordnung der eingeschlossenen Fälle zu Trainings- und Testdatensätzen im Verhältnis 80:20 (siehe Kapitel 3.3.4).

Im **zweiten Schritt (Modellentwicklung und -vergleich**) werden sämtliche Analysetechniken zur Entwicklung und vergleichenden Testung der Regressions- und ML-Modelle durchgeführt. Bei der Auswahl geeigneter ML-Verfahren fiel die Entscheidung zugunsten Random Forest (RF), Adaptive Boosting (AdaBoost) und Künstliche Neuronale Netze (KNN). Die Entwicklung dieser Verfahren wurde jeweils auf den Trainingsdaten aus dem Jahr 2018 vorgenommen. Die Begrenzung auf das Datenjahr 2018 hat den Vorteil, dass für die entsprechenden Fälle nicht nur ein vollständiges Jahr zur Vor- und Nachbeobachtung (sprich 2017 und 2019) vorlag, sondern auch genügend zeitlicher Abstand bis zum Beginn der COVID-19-Pandemie im Jahr 2020 bestand. Das Training der ML-Verfahren erfolgte in mehreren Teilschritten, in denen verfahrensspezifische Methoden der Kreuzvalidierung, des Up- und Downsamplings, der gewichteten Fehlerfunktion sowie der Hyperparameteroptimierung zum Einsatz gelangten (siehe Kapitel 0). Parallel dazu wurden logistische Regressionsmodelle auf den Trainingsdaten 2018 trainiert (siehe Kapitel 3.4.1). Anschließend wurden die trainierten Regressions- und ML-Modelle auf Basis der Testdaten aus dem Jahr 2018 anhand von Evaluationsmetriken geprüft und miteinander verglichen (siehe Kapitel 3.5). Die Modellevaluation wurde zudem ergänzt um weitere Post-hoc-Analysen zur Erklärbarkeit (siehe Kapitel 3.6.1), zu Fehlklassifikationen und zur Übertragbarkeit der Modelle auf die Folgejahre 2019 und 2020 (siehe Kapitel 3.6.2).

3.3 Aufbereitung der GKV-Routinedaten

3.3.1 Grundlegende Aufbereitungsschritte

Da GKV-Routinedaten primär zu Abrechnungszwecken genutzt werden, sind zu Beginn bestimmte Aufbereitungsmaßnamen erforderlich, um die Daten hinsichtlich der projektspezifischen Fragestellungen nutzen zu können. Im Projekt galt dies vor allem für die Krankenhausfalldaten, die einer aufwändigen Aufbereitungsprozedur unterzogen werden mussten, um beispielsweise Mehrfacheinträge zu Krankenhausfällen (z. B. aufgrund von internen Verlegungen auf andere Stationen) anhand bestimmter Kriterien zu realen, abgeschlossenen Krankenhausaufenthalten zusammenzuführen. Dies ist erforderlich, um im weiteren Verlauf valide Berechnungen zu poststationären Ereignissen anstellen zu können. Weitere Aufbereitungsschritte waren unter anderem die Zensur von Fällen aufgrund nicht durchgängiger Versicherungszeiten und die Datenrekodierung zur Reduktion des Rechen- und

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

Speicherbedarfs. Detaillierte Ausführungen zur grundlegenden Aufbereitung der Routinedaten stehen im Kapitel 4.3 des Weißbuchs (siehe Anlage 1) beschrieben.

3.3.2 Operationalisierung der Outcomes

Die Definition und das Vorgehen zur Berechnung der beiden Outcomes Mortalität und Ungeplante Wiederaufnahme (siehe Tabelle 3) wurden aus den Vorläuferprojekten EMSE und USER übernommen. Beim Outcome Ungeplante Wiederaufnahme wurde der Aufnahmegrund „Notfall“ (gem. Anlage 2 zur § 301-Datenübermittlungsvereinbarung, Schlüssel 1) herangezogen, welcher zwar von Krankenhäusern nicht nur bei medizinischen Notfällen kodiert wird, aber erfahrungsgemäß geeignet ist, um zwischen elektiven und nicht-elektiven Krankenhausaufnahmen unterscheiden zu können. Aus diesem Grund wurde sich im Projekt auch auf die Begrifflichkeit „Ungeplante Wiederaufnahme“ anstatt „Notfallwiederaufnahme“ verständigt.

Tabelle 3: Definition der Outcomes

Outcome	Definition	Berechnung
Mortalität	Outcome: Versterben innerhalb von 30 Tagen nach einer vollstationären Krankenhausentlassung (0/1-codiert)	Indexfall wird mit Wert=1 codiert, sobald Todesdatum (sofern vorhanden) zwischen Index-Entlassdatum und Index-Entlassdatum + 30 Tage liegt
Ungeplante Wiederaufnahme	Outcome: Ungeplante Wiederaufnahme ins Krankenhaus innerhalb von 30 Tagen nach einer vollstationären Krankenhausentlassung (0/1-codiert)	Indexfall wird mit Wert=1 codiert, sobald ein Aufnahmedatum ins Krankenhaus mit Aufnahmegrund „Notfall“ zwischen Index-Entlassdatum und Index-Entlassdatum + 30 Tage liegt

3.3.3 Auswahl und Aufbereitung der Prädiktoren

Ausgangspunkt für eine erste Prädiktorauswahl (im ML-Kontext auch „Feature Selection“ genannt) waren die Vorarbeiten der vorangegangenen Projekte EMSE und USER. In diesen Projekten wurden anhand von studienbasierten und eigenen empirischen Erkenntnissen Prädiktoren zusammengestellt, mit denen sich das Eintreten bzw. Nicht-Eintreten der Outcomes Mortalität und Ungeplante Wiederaufnahme vorhersagen lässt. Dabei wurden die Prädiktoren in drei Blöcke eingeteilt. Der erste Block beinhaltet die sogenannten Basisprädiktoren, die eine allgemeine bzw. outcome-unspezifische Erklärungskraft aufweisen. Der zweite Block umfasst sämtliche ambulanten und stationären Diagnosen (binär kodiert auf Ebene der 241 ICD-Gruppen), die im Zeitraum von 365 Tagen vor Krankenhausaufnahme gestellt worden sind. Ergänzend dazu wurde im Projekt KI-THRUST ein dritter Block gebildet, der die „Vorerkrankungsdiagnosen“ des zweiten Blockes nach den Diagnosequartalen ausdifferenziert, um die Relevanz des zeitlichen Abstandes bis zur Aufnahme analytisch zu berücksichtigen. Sämtliche Prädiktoren mit den zugrundeliegenden Definitionen sowie der Zuweisung zu den Blöcken und den finalen Datenmodellen (M1 bis M3) sind der Tabelle 4 zu entnehmen.

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

Tabelle 4: Übersicht der Prädiktorvariablen und Datenmodelle

Prädiktorvariable	Beschreibung	
Datenmodell 1 (M1)		
Alter	Alter in Jahren zum Zeitpunkt der Aufnahme	Basisprädiktoren
Geschlecht	Geschlecht (männlich/weiblich: 0/1-codiert)	
Mehrfacher Krankenhausaufenthalt	Mehr als ein Krankenhausaufenthalt innerhalb von 6 Monaten vor der Aufnahme (0/1-codiert)	
Langer Krankenhausaufenthalt	Mindestens ein Krankenhausaufenthalt mit Verweildauer >21 Tagen in den 365 Tagen vor Aufnahme (0/1-codiert)	
Polymedikation	Mindestens 6 unterschiedliche Arzneimittelverordnungen innerhalb von 3 Monaten vor Aufnahme (0/1-codiert)	
Hilfsmittel	Mindestens eine Hilfsmittelverordnung in den 365 Tagen vor Aufnahme (0/1-codiert)	
Pflegegrad	Pflegegrad zum Zeitpunkt der Aufnahme (von 0 bis 5)	
Datenmodell 2 (M2)		
Basisprädiktoren	s. Modell M1	
ICD: A00-A09 bis ICD: Z80-Z99	Vorhandensein min. einer ICD-Diagnose (ambulant oder stationär) innerhalb der jeweiligen dreistelligen ICD-Gruppe (insgesamt 241 verschiedene ICD-Gruppen: A00-A09, A15-A19, A20-A28 … usw. bis Z80-Z99, somit 241 unterschiedliche Variablen, jeweils 0/1-codiert) in den 365 Tagen vor Aufnahme ins Krankenhaus	
Datenmodell 3 (M3)		
Basisprädiktoren	s. Modell M1	
ICD: A00-A09(Q1), ICD: A00-A09(Q2), ICD: A00-A09(Q3), … bis ICD: Z80-Z99(Q4)	Vorhandensein min. einer ICD-Diagnose (ambulant oder stationär) der ICD-Gruppe im jeweiligen 3 Monatszeitraum von Q1 (1.-3. Monat vor KH-Aufnahme) bis Q4 (9.-12. Monat vor KH-Aufnahme); insgesamt 4 x 241 Variablen, jeweils 0/1-codiert)	

3.3.4 Datenselektion und Trainings-Test-Split

Wie aus der Abbildung 2 hervorgeht, waren in den rohen Krankenhausfalldaten insgesamt 4.090.658 Krankenhausfälle enthalten. Aus diesen konnten im Rahmen der Daten- und Intervallaufbereitung 3.655.748 abgeschlossene, vollstationäre Krankenhausepisoden (mit einer Entlassung nach Hause oder in eine Nicht-Krankenhaus-Einrichtung) identifiziert und zusammengeführt werden. Von den weiterführenden Analysen ausgeschlossen wurden 234.232 (5,7 %) Geburtsfälle, 5.718 (0,16 %) Kinder im Alter von unter einem Jahr sowie 776.635 (21,2 %) Fälle aufgrund nicht durchgängiger Versicherungszeiten.

Im nächsten Schritt wurden die für die Analysen eingeschlossenen Krankenhausfälle in Trainingsdaten (n=2.297.837) und Testdaten (n=575.558) im Verhältnis 80 % zu 20 % aufgeteilt. Die Zuweisung erfolgte über eine Randomisierung der Versichertenpseudonyme. Anschließend wurden die Test- und Trainingsdaten nach Jahresscheiben gesplittet. Das Modelltraining basierte auf den Trainingsdaten mit einer Entlassung im Jahr 2018. Für die anschließende Testung der Vorhersagemodelle wurden neben dem Datenjahr 2018 auch die

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

Datenjahre 2019 und 2020 verwendet, wobei für das Jahr 2020 nur vollständige Daten bis zum 30. September zur Verfügung standen.

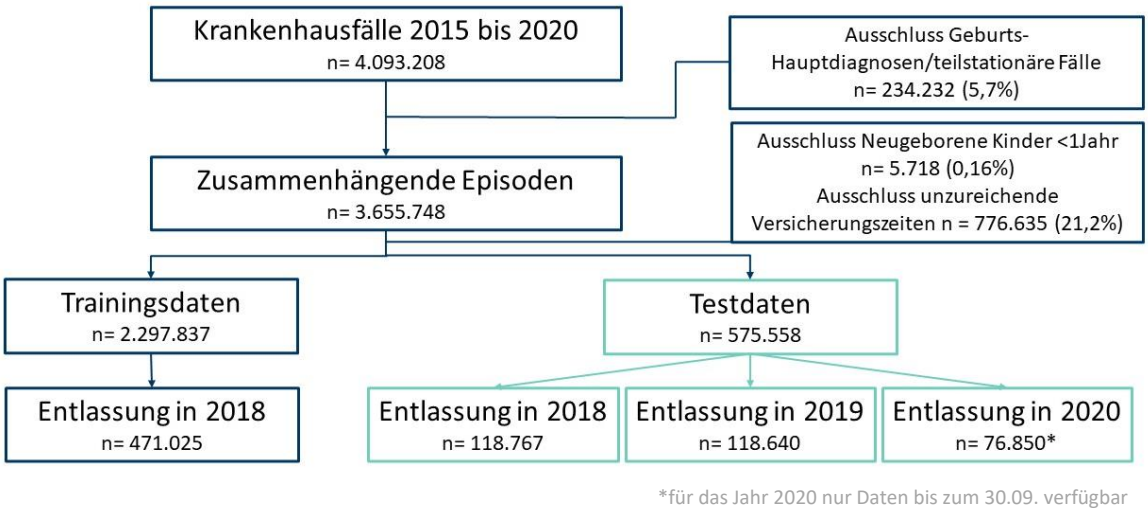

Abbildung 2: Flussdiagramm zur Fallselektion und Aufteilung in Test- und Trainingsdaten

3.4 Modelltraining

3.4.1 Logistisches Regressionsmodell

Als Stellvertreter für ein konventionelles statistisches Verfahren wurden multivariate logistische Regressionsmodelle zur Vorhersage der binär kodierten Outcomes Mortalität und Ungeplante Wiederaufnahme berechnet. Die Modellentwicklung erfolgte, wie bei den Verfahren des Maschinellen Lernens, auf Basis der Trainingsdaten aus dem Jahr 2018. Anschließend wurden die entwickelten Modelle auf den Testdaten 2018 sowie zur Prüfung der Übertragbarkeit auf den Testdaten 2019 und 2020 evaluiert. Sämtliche Berechnungen wurden mit der Statistiksoftware SAS und entsprechenden SAS-Prozeduren (v. a. proc logistic) auf speziellen SAS-Servern des aQua-Instituts durchgeführt. Insgesamt wurden je Outcome drei Modelle aufgestellt, deren Modellformulierung den im Kapitel 3.3.3 beschriebenen Datenmodellen M1 bis M3 entspricht.

Das **erste Modell (M1)** umfasst die sogenannten Basisprädiktoren (siehe Tabelle 4). Die darin enthaltenen kategorialen Variablen (z. B. Altersgruppen) wurden im Zuge der Regressionsanalyse mittels einer Dummy-Variablenkodierung und einer entsprechenden Festlegung einer Referenzkategorie (z. B. Altersgruppe 60-64 Jahre) rekodiert. Alle Basisprädiktoren wurden als Block per Einschlussverfahren in das Modell aufgenommen.

Das **zweite Modell (M2)** enthielt neben den Basisprädiktoren zusätzlich die ambulanten und stationären Vorerkrankungsdiagnosen (siehe Tabelle 4). Die Basisprädiktoren wurden im ersten Block per Einschluss aufgenommen. Der zweite Block bestand aus dem Pool aller 241 Vorerkrankungsdiagnosen, aus denen die statistisch relevantesten Diagnosen per Vorwärtsselektion (stepwise-forward-selection) ausgewählt wurden. Dabei wurde das Modell schrittweise um Variablen mit dem jeweils größten F-Wert erweitert, sofern der p-Wert unter 0,05 lag. Bereits in der Gleichung enthaltene Variablen wurden ausgeschlossen, sobald der p-Wert über 0,1 stieg.

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

Im **dritten Modell (M3)** waren neben den Basisprädiktoren die zeitlich differenzierten Vorerkrankungsdiagnosen enthalten (siehe Tabelle 4). Wie bei Modell M2 wurden im Modell M3 zunächst die Basisprädiktoren in einem ersten Block in das Modell aufgenommen. Danach wurden als zweiten Block die quartalsbezogenen Vorerkrankungsdiagnosen (4 x 241 Variablen) iterativ durch eine Vorwärtsselektion (stepwise-forward-selection) mit denselben Selektionskriterien wie bei Modell M2 eingeschlossen.

3.4.2 Machine Learning-Verfahren

Im Projekt erfolgte die Modellierung auf Grundlage etablierter Verfahren des Maschinellen Lernens. Das Ziel bestand darin, leistungsstarke Modelle zur Prädiktion der Outcomes Mortalität und Ungeplante Wiederaufnahme zu entwickeln und zu bewerten. Bei der Auswahl der Verfahren wurde auf eine Vielfalt unterschiedlicher Modellklassen geachtet, die sowohl lineare als auch nicht lineare Zusammenhänge erfassen können.

Für alle Modellierungsansätze wurde eine umfassende Datenvorverarbeitung durchgeführt. Hierzu zählte insbesondere die Transformation kategorialer Merkmale durch das One-Hot-Encoding, um eine binäre Repräsentation sämtlicher Ausprägungen zu gewährleisten. Die vollständige Kodierung ohne Reduktion auf eine Referenzkategorie wurde beibehalten, um potenzielle Informationsverluste zu vermeiden und um auch den Entscheidungsbaumverfahren eine differenzierte Berücksichtigung aller Merkmalsausprägungen zu ermöglichen.

Zur Vorhersage der Zielgrößen kamen drei Modelle zum Einsatz: ein AdaBoost-Klassifikator, ein Random-Forest-Klassifikator und ein tiefes Künstliches Neuronales Netz (KNN) in Form eines Multilayer Perceptrons (MLP). AdaBoost ist ein sequenzielles Ensemble-Verfahren, das auf der iterativen Kombination “schwacher” Klassifikatoren basiert. Dabei wird in jeder Iteration die Gewichtung der Trainingsbeispiele angepasst, um Fehlklassifikationen gezielt zu korrigieren. Dieses Vorgehen ermöglicht eine schrittweise Verbesserung der Modellleistung und eignet sich besonders bei nichtlinearen Entscheidungsgrenzen und heterogenen Datenverteilungen.

Random Forest kombiniert hingegen eine Vielzahl unabhängiger Entscheidungsbäume, wobei jeder Baum auf zufälligen Teilmengen der Daten und Features trainiert wird („Bagging“ + Merkmal-Randomisierung). Diese Modellklasse zeichnet sich durch eine hohe Robustheit gegenüber Überanpassung („Overfitting“) und Ausreißern sowie durch eine gute Interpretierbarkeit der Merkmalsbedeutungen aus. Gerade bei zahlreichen kategorischen Merkmalen liefern baumbasierte Modelle tendenziell stabilere und leichter erklärbare Ergebnisse als Künstliche Neuronale Netze, da Entscheidungsbäume intuitiv interpretierbare Splits auf nominalen Kategorien erlauben.

Das MLP erlaubt durch seine mehrschichtige, vollständig verbundene Struktur die Abbildung komplexer, nichtlinearer Zusammenhänge. Es bietet die größte Modellflexibilität und Skalierbarkeit gegenüber größeren Datensätzen, sodass es Muster erfassen kann, die Entscheidungsbaum-basierte Modelle eventuell nicht abbilden. Allerdings erfordert das MLP eine sorgfältige Regularisierung (z.B. Dropout) und Hyperparameter-Optimierung, um Overfitting zu vermeiden.

Angesichts der stark unausgeglichenen Klassenverteilung der Zielgrößen, insbesondere der Mortalität, wurden verschiedene Strategien zur Korrektur dieser Unbalanciertheit

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

implementiert. Einerseits erfolgte ein synthetisches Oversampling der Minderheitsklasse. Dadurch konnten realistische zusätzliche Instanzen erzeugt werden, die zur Stabilisierung des Lernprozesses beitrugen. Ergänzend hierzu wurde ein Downsampling der Mehrheitsklasse vorgenommen, bei dem ein Großteil der überrepräsentierten Instanzen entfernt wurde, um ein annähernd ausgewogenes Klassenverhältnis herzustellen. Als zweite Maßnahme wurde eine gewichtete Fehlerfunktion eingeführt, bei der Fehlklassifikationen der Minderheitsklasse im Lernprozess stärker gewichtet wurden. Diese Technik erlaubt eine balancierte Optimierung, ohne dass die Fallzahlen explizit verändert werden müssen. Je nach Modelltyp wurden die genannten Verfahren in unterschiedlicher Kombination eingesetzt. Details zur jeweiligen Modellvariante finden sich im Weißbuch (siehe Anlage 1, Kapitel 5.4).

Die Auswahl der geeigneten Modellparameter erfolgte durch eine systematische Hyperparameteroptimierung mittels Gittersuche (engl. grid search) in Verbindung mit einer 5-fachen Kreuzvalidierung. Um eine verzerrungsfreie Einschätzung der Modellgüte bei starker Klassenunbalanciertheit zu gewährleisten, wurde die Fläche unter der Precision-Recall-Kurve (AUC-PR) als zentrales Optimierungskriterium verwendet. Die getesteten Parameterkombinationen umfassten unter anderem die Anzahl der Basis-Klassifikatoren und Lernraten für AdaBoost, die Tiefe und Anzahl der Entscheidungsbäume sowie das Trennkriterium für Random Forest und verschiedene Lernraten und Batch-Größen für das MLP. Die vollständige Auflistung der getesteten Kombinationen ist im Weißbuch dokumentiert (vgl. Anlage 1, Tabelle 53).

Auf Basis der Ergebnisse der ersten Gittersuche (Modell M1) wurde entschieden, in den weiteren Modellen (M2 und M3) ausschließlich die Modelle mit gewichteter Fehlerfunktion weiter zu optimieren. Die Analyse hatte gezeigt, dass Upsampling und Downsampling vergleichbare Leistungsverbesserungen erzielten, jedoch mit höherem Rechenaufwand und potenziellem Informationsverlust einhergingen. Die Konzentration auf gewichtete Fehlerfunktionen erlaubte somit eine effiziente und robuste Optimierung der Modellleistung bei gleichzeitig geringerer Komplexität.

3.5 Modellevaluation I: Testung und Vergleich

Ziel der Evaluation war es, die prognostische Leistungsfähigkeit der komplexen ML-Modelle im Vergleich zum etablierten Regressionsverfahren quantifizierbar zu machen. Für die Bewertung wurde ein dedizierter Testdatensatz verwendet, der Daten aus dem Zeitraum 2018 bis 2020 enthielt. Während der gesamten Phase der Modellentwicklung und Hyperparameteroptimierung blieben die Testdaten vollständig unberührt. Dadurch wurde eine realistische und unverzerrte Einschätzung der Generalisierungsfähigkeit der Modelle ermöglicht.

Zur Leistungsbewertung kamen zwei etablierte Evaluationsmetriken zum Einsatz: die Fläche unter der Receiver Operating Characteristic-Kurve (Area Under the Receiver Operating Characteristic Curve, AUC-ROC) sowie die Fläche unter der Precision-Recall-Kurve (Area Under the Precision-Recall Curve, AUC-PR). Die AUC-ROC beschreibt die Fähigkeit eines Modells, zwischen den Klassen zu unterscheiden, unabhängig von einem spezifischen Schwellenwert, und bietet eine Gesamtbetrachtung der Trade-offs zwischen Sensitivität (Richtig-Positiv-Rate) und 1-Spezifität (Falsch-Positiv-Rate).

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Die AUC-PR hingegen legt den Fokus stärker auf das Verhältnis von Präzision (Positiv-Prädiktiver-Wert) und Sensitivität (Richtig-Positiv-Rate). Diese Metrik ist insbesondere bei stark unausgeglichenen Klassenverteilungen von Bedeutung, wie sie bei seltenen Ereignissen (z. B. Mortalität) auftreten. In solchen Szenarien erlaubt sie eine differenziertere und praxisnähere Bewertung der Modellleistung.

Eine ausführliche Beschreibung der verwendeten sowie weiterer Evaluationsmetriken findet sich im Weißbuch (siehe Anlage 1, Kapitel 3).

3.6 Modellevaluation II: Ergänzende Post-hoc-Analysen

3.6.1 Erklärbarkeit

Ein zentrales Anliegen des Projekts war es, die Entscheidungen datengetriebener Modelle nicht nur quantitativ zu bewerten, sondern auch qualitativ nachvollziehbar zu machen. Gerade in sensiblen medizinischen Bereichen wie der Vorhersage von Mortalität oder ungeplanten Krankenhauswiederaufnahmen ist die Interpretierbarkeit von Vorhersagen von entscheidender Bedeutung. Nur so kann Vertrauen in automatisierte Systeme geschaffen werden und potenzielle Verzerrungen oder fehlerhafte Lernmechanismen können frühzeitig erkannt werden. Aus diesem Grund wurde für alle trainierten Modelle eine systematische Analyse der Modellentscheidungen durchgeführt, bei der sowohl intrinsische als auch post-hoc-Erklärbarkeitsansätze berücksichtigt wurden.

Für die beiden Ensemble-Modelle AdaBoost und Random Forest konnten intrinsische Merkmalsbedeutungen genutzt werden. Diese basieren auf der mittleren Reduktion eines Qualitätskriteriums, typischerweise der Gini-Impurity, durch Splits entlang eines Merkmals. Die summierten Reduktionsbeiträge über alle Entscheidungsbäume hinweg liefern eine gewichtete Importance-Metrik, die als globales Maß für die Relevanz einzelner Merkmale interpretiert werden kann. Dabei ist jedoch zu beachten, dass dieser Ansatz methodenbedingt systematische Verzerrungen aufweisen kann, etwa zugunsten kontinuierlicher Merkmale oder solcher mit vielen Kategorien. Diese Limitierung wurde bei der Interpretation der Ergebnisse berücksichtigt.

Um die Entscheidungslogik auch bei Black-Box-Modellen nachvollziehbar zu machen, wurden für das MLP sowie ergänzend auch für die beiden anderen Modellklassen Post-hoc-Erklärungsverfahren angewendet. Hierzu wurden drei unterschiedliche Methoden eingesetzt: Integrated Gradients, LIME und Shapley Value Sampling. Bei der Auswahl dieser Verfahren wurde auf ihre theoretische Fundierung, ihre Modellunabhängigkeit sowie ihre Etablierung in der wissenschaftlichen Literatur geachtet.

Integrated Gradients wurde speziell für Neuronale Netze entwickelt. Die Methode berechnet für jede Eingabevariable die Veränderung der Modellvorhersage entlang eines linearen Pfads zwischen einem Baseline-Input und dem tatsächlichen Input. Die Methode liefert differenzierte Attributionswerte und ist insbesondere für tiefere Netzarchitekturen geeignet. LIME basiert dagegen auf der lokalen Approximation von Modellentscheidungen durch einfach interpretierbare Surrogatmodelle. Diese Methode eignet sich besonders für die Analyse einzelner Fälle.

Shapley Values stellen schließlich ein spieltheoretisch fundiertes Verfahren dar, bei dem der durchschnittliche marginale Beitrag eines Merkmals über alle möglichen Kombinationen

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

anderer Merkmale hinweg berechnet wird. Die Methode ist sowohl global als auch lokal interpretierbar und erlaubt eine faire Verteilung der Relevanz über alle Merkmale.

Zur methodenübergreifenden Vergleichbarkeit wurden alle Relevanzwerte normalisiert. Hierzu wurde eine Rangnormalisierung durchgeführt, bei der die Relevanzen komponentenweise geordnet und anschließend skaliert wurden. Diese Vorgehensweise reduziert methodenspezifische Skaleneffekte und ermöglicht eine einheitliche Bewertung der Bedeutung einzelner Merkmale sowohl innerhalb eines Modells als auch zwischen unterschiedlichen Modellklassen.

3.6.2 Übertragbarkeit und Fehlklassifikationen

Um die Übertragbarkeit der auf dem Datenjahr 2018 trainierten Modelle auf nachfolgende bzw. „zukünftige“ Jahre zu untersuchen, wurden das beste logistische Regressionsmodell sowie das beste ML-Modell genutzt, um die Prognosegüte für die beiden Outcomes mit Daten aus den Jahren 2019 und 2020 zu testen und zu vergleichen. Für die logistische Regression kam das Datenmodell M2 zum Einsatz. Als Machine Learning-Verfahren wurde AdaBoost (ebenfalls Datenmodell M2) zu Vergleichszwecken herangezogen. Die Daten aus den Jahren 2019 und insbesondere 2020 unterscheiden sich in mehreren Aspekten signifikant vom Jahr 2018. Um Unterschiede in der Prognosegüte zu identifizieren, wurden die beiden Evaluationsmetriken AUC-ROC und AUC-PR, also die Flächen unter den Kurven, für alle Datenjahre berechnet und miteinander verglichen (detaillierte Beschreibung s. a. Weißbuch, Anlage 1, Kapitel 8.3.1).

Um die Prognosegüte der Modelle in unterschiedlichen Subgruppen zu untersuchen, wurden das beste logistische Regressionsmodell sowie das AdaBoost-Verfahren als bestes ML-Modell (jeweils Datenmodell M2) genutzt, um für jede Subgruppe die Receiver Operating Characteristic (ROC-Kurve) zu berechnen und den Wert für die Fläche unter der Kurve (AUC-ROC) mit Angabe des 95 %-Konfidenzintervalls miteinander zu vergleichen. Für die Variablen Geschlecht (männlich/weiblich) und Alter (in 20-Jahres-Altersgruppen) wurden Subgruppen gebildet, sowie für die Art des Krankenhausaufenthalts des Indexfalls (Normalfall vs. Notfall). Um die Auswirkung der Datenverfügbarkeit zu untersuchen, wurden Patienten/Patientinnen aus Pflegeheimen untersucht. Dazu wurden Patienten/Patientinnen, die im Jahr vor Aufnahme ins Krankenhaus mindestens 90 Tage in stationärer Pflege waren, verglichen mit Patienten/Patientinnen, die gar nicht oder kürzer in stationärer Pflege waren (s. a. Anlage 1, Kapitel 8.3.2).

4 Projektergebnisse

4.1 Deskription der Versichertenpopulation

Die Versicherten im Trainingsdatensatz waren im Mittel 61,2 Jahre alt (Range von 1 bis 109 Jahre, SD = 22,85). Der Anteil an Männern im Datensatz lag bei 49,7 %, der Anteil Frauen bei 50,3 %. 82 Versicherte hatten den Geschlechtseintrag „divers“ oder „unbekannt“. Diese Versicherten wurden aufgrund der geringen Anzahl vor der Erstellung des Trainingsdatensatzes ausgeschlossen. Eine detaillierte Beschreibung der Testdatensätze sowie Angaben zur Häufigkeit der untersuchten Outcomes und der Basisprädiktoren finden sich in Tabelle 5. Für die in der Tabelle beschriebenen Kennwerte zeigte sich kein signifikanter

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Unterschied zwischen dem Trainingsdatensatz (aus dem Jahr 2018) und dem Testdatensatz 2018 (alle $p > 0,17$), was der Intention bei einer zufälligen Aufteilung in Trainings- und Testdaten entspricht. Die verwendeten Testdatensätze aus den Jahren 2019 und 2020 enthielten demgegenüber signifikant mehr Versicherte mit Pflegegrad und Hilfsmittelverordnungen (alle $p < 0,01$). Im Testdatensatz 2020 kam außerdem das Outcome Mortalität signifikant häufiger vor ($p < 0,01$) und die Versicherten waren im Mittel 0,5 Jahre älter als im Trainingsdatensatz 2018 ($p < 0,01$), wobei sich die Unterschiede in einem Rahmen bewegen, der bei Daten zu Populationen aus unterschiedlichen Beobachtungsjahren zu erwarten ist.

Tabelle 5: Beschreibung der Versichertenpopulation nach ausgewählten Merkmalen in Test- und Trainingsdaten

Merkmal	Datensatz			
	Trainingsdaten 2018	Testdaten 2018	Testdaten 2019	Testdaten 2020
Stichprobengröße (n)	471.025	118.767	118.640	76.850*
Alter (M ± SD)	61,2 ± 22,9	61,2 ± 22,9	61,3 ± 22,9	61,7 ± 22,8
Altersgruppen in Jahren (%):				
Unter 20	6,7	6,8	6,7	6,7
20 bis 39	12,2	12,2	12,3	12,2
40 bis 59	20,3	20,3	20,0	19,5
60 bis 79	36,7	36,7	36,2	35,7
80 und älter	24,0	24,1	24,9	26,2
Geschlecht (%):				
Weiblich	50,3	50,5	50,2	50,2
Männlich	49,7	49,5	49,8	49,8
Pflegegrad (%):				
0	85,6	85,6	83,1	73,9
1	1,0	1,1	1,4	2,4
2	6,0	6,0	6,5	9,4
3	4,3	4,3	5,2	8,1
4	2,2	2,3	2,8	4,6
5	0,9	0,8	1,0	1,7
Outcome „Mortalität“ (%)	1,0	1,0	1,0	1,2
Outcome „Ungeplante Wiederaufnahmen“ (%)	8,2	8,2	8,3	7,9
Prädiktor „Polymedikation“ (%)	38,0	38,0	38,2	37,9
Prädiktor „Mehrfache KH-Aufenthalte“ (%)	16,0	16,2	16,4	15,6
Prädiktor „Langer KH-Aufenthalt“ (%)	8,8	8,9	9,0	8,7
Prädiktor „Hilfsmittelbedarf“ (%)	38,2	38,2	45,0	51,6

*für das Jahr 2020 wurden nur Daten bis zum 30.09. berücksichtigt

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

4.2 Modellevaluation I: Modelltestung und -vergleich

Nach der Architektur- und Parameteroptimierung der verschiedenen ML-Modelle anhand der Trainingsdaten aus dem Jahr 2018 wurden jeweils die Modelle mit der besten Variante und Leistung für den Vergleich mit der logistischen Regression ausgewählt. Der Modellvergleich anhand der Metriken AUC-ROC und AUC-PR wurde mit Testdaten aus dem Jahr 2018 durchgeführt.

Outcome Mortalität

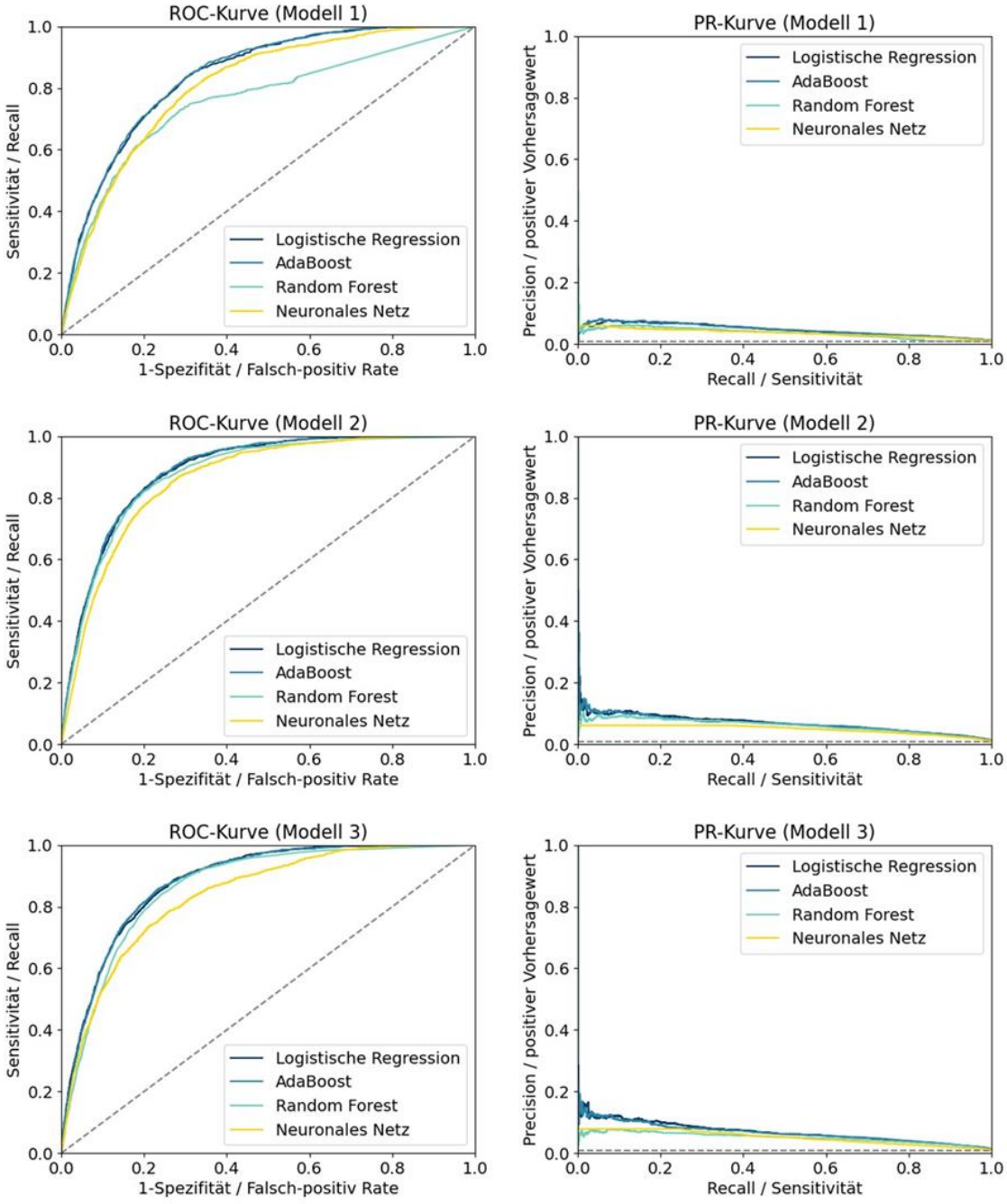

*Abbildung 3: Vergleich der ML- und Regressionsmodelle zur Vorhersage der **Mortalität** anhand von ROC- und PR-Kurven mit Testdaten aus dem Jahr 2018*

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Die Bewertung der Modelle zur Vorhersage des Outcomes Mortalität erfolgte anhand der ROC-Kurven (siehe Abbildung 3) und der zugehörigen AUC-ROC-Werte (siehe Tabelle 6). Über alle Klassifikatoren hinweg zeigte sich ein deutlicher Leistungszuwachs von Modell M1 zu M2. Die Integration von Vorerkrankungsdiagnosen in M2 erhöhte die Trennschärfe signifikant gegenüber dem Basismodell, das ausschließlich die Basisprädiktoren berücksichtigt. Eine zusätzliche zeitliche Differenzierung dieser Diagnosen in Modell M3 führte hingegen zu keiner weiteren Verbesserung, sondern zu leichten Einbußen in der Performance. Auf Grundlage des Datenmodells M2 wurden konsistent die besten Resultate über alle Verfahren hinweg erzielt.

Die AUC-ROC-Werte von Modell M2 lagen bei allen vier Verfahren im Bereich von 0,861 bis 0,889 und wurden gemäß gängiger Schwellenwerte als „ausgezeichnet“ bewertet. AdaBoost (0,889) und die logistische Regression (0,888) zeigten die besten Resultate, gefolgt von Random Forest (0,878) und dem Neuronalen Netz (0,861). Besonders auffällig war die schwache Performance von Random Forest im Basismodell M1, die jedoch durch die Aufnahme zusätzlicher Prädiktoren in M2 und M3 kompensiert werden konnte.

*Tabelle 6: Modellgüte der Regressions- und ML-Modelle zur Vorhersage der **Mortalität** mit Testdaten aus dem Jahr 2018*

Outcome: Mortalität	Modell 1		Modell 2		Modell 3	
Verfahren	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR
Logistische Regression	0,838**	0,046	0,888**	0,068	0,882**	0,069
AdaBoost	0,841**	0,048	0,889**	0,068	0,882**	0,068
Random Forest	0,755*	0,036	0,878**	0,061	0,864**	0,051
Neuronales Netz	0,807**	0,037	0,861**	0,048	0,838**	0,051

*akzeptabel (0,7 ≤ AUC-ROC < 0,8) **ausgezeichnet (0,8 ≤ AUC-ROC < 0,9) ***hervorragend (0,9 ≤ AUC-ROC)

Die PR-Analyse bestätigte den Leistungszuwachs von M1 zu M2 auch hinsichtlich der AUC-PR-Werte. In Modell M3 war jedoch ein Rückgang bei Random Forest und dem Künstlichen Neuronalen Netz zu verzeichnen, während AdaBoost stabil blieb und die logistische Regression eine leichte Verbesserung aufwies (von 0,068 auf 0,069). Trotz AUC-PR-Werten oberhalb des Random-Baseline-Niveaus (Nullmodell, 0,0098) blieb die absolute Präzision insgesamt niedrig. Dies weist auf eine begrenzte positive Vorhersagekraft hin, selbst bei hoher Trennschärfe, was bei der Anwendung unter stark unausgeglichenen Klassenverteilungen berücksichtigt werden muss.

Outcome Ungeplante Wiederaufnahmen

Die Modellgüte zur Vorhersage ungeplanter Wiederaufnahmen wurde – analog zur Evaluation des Outcomes Mortalität – anhand der ROC-Kurven (siehe Abbildung 4) und den zugehörigen AUC-Werten (siehe Tabelle 7) systematisch analysiert. Über alle eingesetzten Klassifikatoren hinweg zeigte Modell M2 die besten Ergebnisse, was die Relevanz der Integration von Vorerkrankungsdiagnosen als zusätzliche Merkmale zu den Basisprädiktoren (Alter, Geschlecht etc.) unterstreicht. Dennoch lagen alle AUC-ROC-Werte unterhalb von 0,7, der gängigen Schwelle für eine „akzeptable“ Diskriminationsleistung. Dies deutet auf eine deutlich geringere Modellierbarkeit dieses Outcomes hin, insbesondere im Vergleich zur Mortalitätsprognose, und verweist auf strukturelle Limitationen der zugrundeliegenden Datenbasis.

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

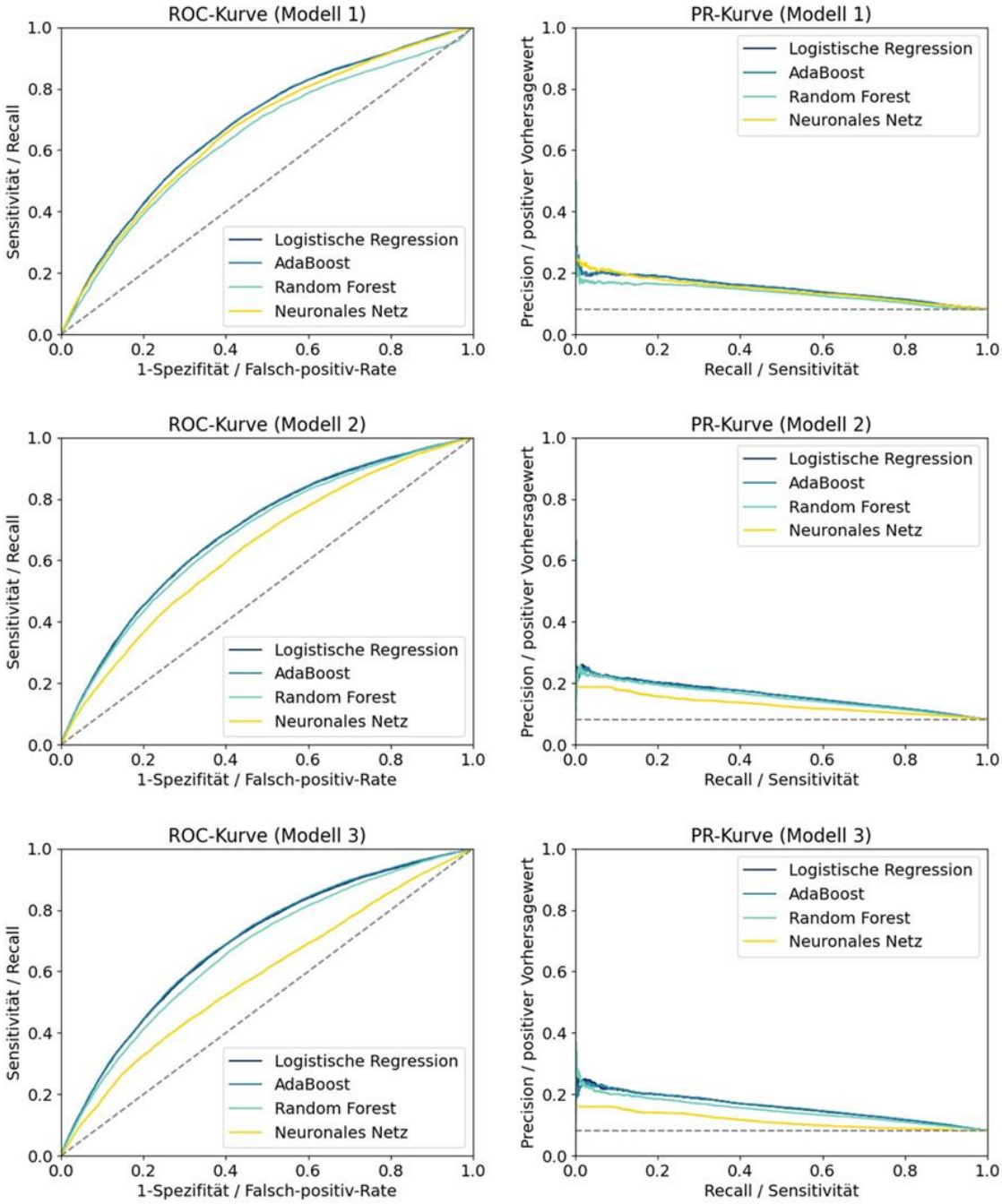

*Abbildung 4: Vergleich der ML- und Regressionsmodelle zur Vorhersage der **Ungeplanten Wiederaufnahmen** anhand von ROC- und PR-Kurven mit Testdaten aus dem Jahr 2018*

Beim Vergleich der Klassifikatoren (jeweils mit Modell M2) ergab sich eine klare Rangfolge: AdaBoost (AUC-ROC = 0,694) und logistische Regression (0,693) lieferten die höchsten Trennschärfen, gefolgt von Random Forest (0,681) und dem Künstlichen Neuronalen Netz (0,638). Auffällig war, dass das Neuronale Netz im Basismodell (M1) eine vergleichsweise gute Leistung zeigte, mit jedoch deutlich abfallender Performance in den Modellen M2 und M3. Dieses Verhalten legt nahe, dass das Netz weniger effektiv von der Erweiterung um strukturelle Diagnosemerkmale profitiert, möglicherweise aufgrund unzureichender Modellkapazität oder fehlender Regularisierung bei hochdimensionalen Eingaben.

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

*Tabelle 7: Modellgüte der Regressions- und ML-Modelle zur Vorhersage der **Ungeplanten Wiederaufnahmen** mit Testdaten aus dem Jahr 2018*

Outcome: Ungeplante Wiederaufnahmen	Modell 1		Modell 2		Modell 3	
Verfahren	**AUC-ROC**	**AUC-PR**	**AUC-ROC**	**AUC-PR**	**AUC-ROC**	**AUC-PR**
Logistische Regression	0,678	0,150	0,693	0,161	0,690	0,159
AdaBoost	0,678	0,150	0,694	0,160	0,693	0,159
Random Forest	0,641	0,134	0,681	0,154	0,669	0,148
Neuronales Netz	0,665	0,146	0,638	0,131	0,584	0,114

Zusätzliche Validierung erfolgte über Precision-Recall-Kurven (siehe Abbildung 4, rechts) und die zugehörigen AUC-PR-Werte (siehe Tabelle 7). Die Rangfolge der Verfahren entsprach weitgehend den ROC-basierten Resultaten, was auf eine robuste Korrelation beider Metriken hindeutet. Die AUC-PR-Werte bewegten sich im Bereich von 0,13 bis 0,16 und lagen damit nur moderat über dem Referenzwert eines Nullmodells (AUC-PR = 0,08). Dies verweist auf eine insgesamt niedrige positive Vorhersagekraft (PPV), trotz teilweise hoher Diskriminationsfähigkeit. Entsprechend ist selbst innerhalb der vom Modell als Hochrisikogruppen identifizierten Kohorten nur ein kleiner Anteil tatsächlich von einer ungeplanten Wiederaufnahme betroffen – ein kritischer Aspekt bei der operationalen Nutzung solcher Modelle im Versorgungskontext.

4.3 Modellevaluation II: Ergänzende Post-hoc-Analysen

4.3.1 Erklärbarkeit

Im Folgenden werden die Ergebnisse zur Erklärbarkeit für die verschiedenen Modelle des Maschinellen Lernens dargestellt. Sämtliche Berechnungen basieren auf dem zugrundeliegenden Datenmodell M2. Für die interpretierbaren Modelle AdaBoost und Random Forest werden die globalen Feature Importance-Werte dargestellt. Im Gegensatz dazu stehen beim Künstlichen Neuronalen Netz die lokalen XAI-Relevanzzuweisungen, die mithilfe der Methoden Integrated Gradients, Shapley Value Sampling und LIME ermittelt wurden.

Outcome Mortalität

In Abbildung 5 ist erkennbar, dass für die Modelle AdaBoost und Random Forest das Merkmal “Alter” den größten Einfluss auf die Vorhersage von Mortalität hat. Dies lässt sich jedoch vor allem darauf zurückführen, dass Alter die einzige metrische Variable im Datensatz ist und es die bereits beschriebene systematische Verzerrung gibt. Entsprechend kann aus den reinen Feature-Importance-Werten abgeleitet werden, dass Alter zwar die größte Aufmerksamkeit im Modellierungsprozess erfährt, inhaltlich jedoch nicht die wichtigste Variable für die Vorhersage darstellt.

Alle weiteren Variablen weisen vergleichsweise geringe Feature-Importance-Werte auf (<0,05). Beim AdaBoost-Modell folgen auf Alter die Merkmale „ICD-Gruppe F00-F09“ (Organische, einschließlich symptomatischer psychischer Störungen), „Mehrfache Krankenhausaufenthalte“ (innerhalb von 6 Monaten), sowie „Kein Pflegegrad“ mit den

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

nächsthöheren Werten. Für das Random Forest Modell haben die Merkmale „ICD-Gruppe F00_F09“ (s.o.), „ICD-Gruppe C76-C80“ (Bösartige Neubildungen ungenau bezeichneter, sekundärer und nicht näher bezeichneter Lokalisationen) und „Polymedikation“ neben dem „Alter“ den größten Einfluss.

Die Werte der Erklärbarkeitsmethoden für das Neuronale Netz unterschieden sich stark untereinander. Integrated Gradients hat nur positive Relevanzzuweisungen für „ICD-Gruppe I10-I15“ (Hypertonie) und „ICD-Gruppe Z80-Z99“ (Personen mit potenziellen Gesundheitsrisiken aufgrund der Familien- oder Eigenanamnese). Shapley Value Sampling hat die größte Relevanzzuweisung für „ICD-Gruppe M50-M54“ (Sonstige Krankheiten der Wirbelsäule und des Rückens) gefolgt von „ICD-Gruppe H49-H52“ (Affektionen der Augenmuskeln, Störungen der Blickbewegungen sowie Akkommodationsstörungen und Refraktionsfehler). Gleichermaßen hat „ICD-Gruppe H49-H52“ (s.o.) die größte Relevanzzuweisung für die LIME-Methode, gefolgt von „ICD-Gruppe Z00-Z13“ (Personen, die das Gesundheitswesen zur Untersuchung und Abklärung in Anspruch nehmen).

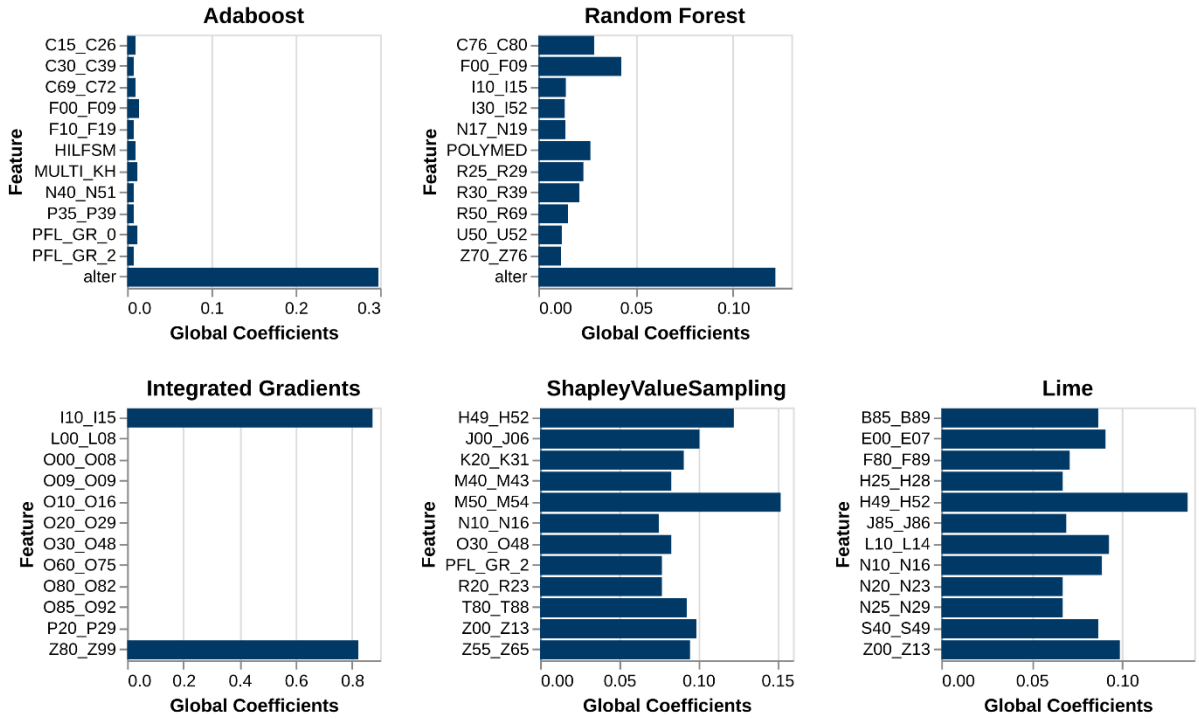

*Abbildung 5: Koeffizienten zur Globale Feature Importance für AdaBoost und Random Forest (oben) sowie zu XAI-Methoden für Neuronales Netz (unten), sortiert nach Top 12, Outcome **Mortalität***

Outcome Ungeplante Wiederaufnahme

Vergleichbar zur Abbildung 5 zeigt auch Abbildung 6, dass das Merkmal „Alter“ in den Modellen AdaBoost und Random Forest den größten Einfluss auf die Vorhersage von Ungeplanter Wiederaufnahme hat. Dieser Befund deckt sich mit den vorherigen Ergebnissen und bestätigt erneut, dass die Dominanz von „Alter“ vor allem auf seine besondere Eigenschaft als einzige metrische Variable im Datensatz zurückzuführen ist. Für beide Modelle folgen „Mehrfache, vorherige Krankenhausaufenthalte“ und „Polymedikation“ mit den nächstgrößten Einflusswerten auf die Vorhersage.

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Die Ergebnisse der Erklärbarkeitsmethoden für das Neuronale Netz unterscheiden sich deutlich: Integrated Gradients weist keinerlei positive Relevanzzuweisungen für die Vorhersage von einer ungeplanten Wiederaufnahme auf. Shapley Value Sampling zeigt die größten Relevanzzuweisungen für „ICD-Gruppe M70-M79“ (Sonstige Krankheiten des Weichteilgewebes), „ICD-Gruppe Z00-Z13“ (Personen, die das Gesundheitswesen zur Untersuchung und Abklärung in Anspruch nehmen), „ICD-Gruppe M40-M43“ (Deformitäten der Wirbelsäule und des Rückens) und „ICD-Gruppe M15-M19“ (Arthrose). Bei LIME wird den Merkmalen „ICD-Gruppe M86-M90“ (Sonstige Osteopathien), „ICD-Gruppe B00-B09“ (Virusinfektionen, die durch Haut- und Schleimhautläsionen gekennzeichnet sind) und „ICD-Gruppe G50-G59“ (Krankheiten von Nerven, Nervenwurzeln und Nervenplexus) die größte Relevanz zugeschrieben.

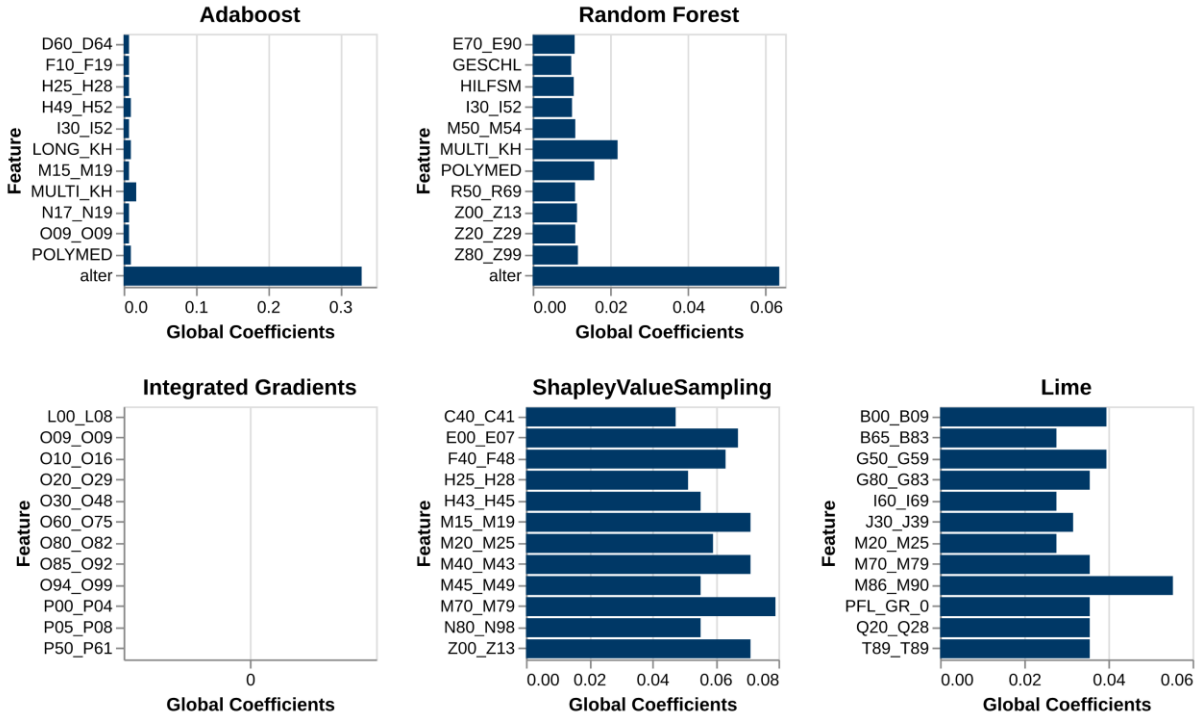

*Abbildung 6: Koeffizienten zur Globale Feature Importance für AdaBoost und Random Forest (oben) sowie zu XAI-Methoden für Neuronales Netz (unten), sortiert nach Top 12, Outcome **Ungeplante Wiederaufnahmen***

4.3.2 Übertragbarkeit und Fehlklassifikationen

Übertragbarkeit auf „zukünftige“ Jahre

Um die Übertragbarkeit der Modelle auf die Jahre 2019 und 2020 zu untersuchen, wurden die auf den Daten des Jahres 2018 entwickelten Modelle auf die jeweiligen Datensätze der Folgejahre angewendet. Die Prognosegüte wurde durch den Vergleich der Flächen unter der ROC-Kurve (AUC-ROC) beurteilt. Die entsprechenden Ergebnisse sind in Tabelle 8 zusammengefasst.

Für das Outcome Mortalität ergaben sich im Jahr 2020 im Vergleich zu 2018 und 2019 etwas niedrigere AUC-ROC-Werte, sowohl für die logistische Regression als auch für das maschinelle Lernverfahren AdaBoost. Hinsichtlich des Outcomes Ungeplante Wiederaufnahmen zeigten sich dagegen über die betrachteten Jahre hinweg keine wesentlichen Unterschiede.

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

*Tabelle 8: Prognosegüte in zukünftigen Jahren für **Mortalität** und **Ungeplante Wiederaufnahmen**, unter Berücksichtigung der Versichertenzahl und Prävalenz*

Verfahren	Datenjahr	AUC-ROC [95 %-KI]	N	Prävalenz Outcome
Outcome: Mortalität (Modell 2)				
Logistische Regression	2018	0,888 [0,880; 0,895]	118.767	0,98 %
	2019	0,893 [0,885; 0,900]	118.640	0,97 %
	2020	0,866 [0,856; 0,876]	76.850	1,15 %
AdaBoost	2018	0,889 [0,882; 0,896]	118.767	0,98 %
	2019	0,892 [0,885; 0,900]	118.640	0,97 %
	2020	0,871 [0,861; 0,881]	76.850	1,15 %
Outcome: Ungeplante Wiederaufnahmen (Modell 2)				
Logistische Regression	2018	0,693 [0,688; 0,698]	118.767	8,18 %
	2019	0,697 [0,691; 0,702]	118.640	8,30 %
	2020	0,692 [0,685; 0,698]	76.850	7,93 %
AdaBoost	2018	0,694 [0,689; 0,699]	118.767	8,18 %
	2019	0,697 [0,692; 0,703]	118.640	8,30 %
	2020	0,692 [0,686; 0,699]	76.850	7,93 %

Prognosegüte in Subgruppen

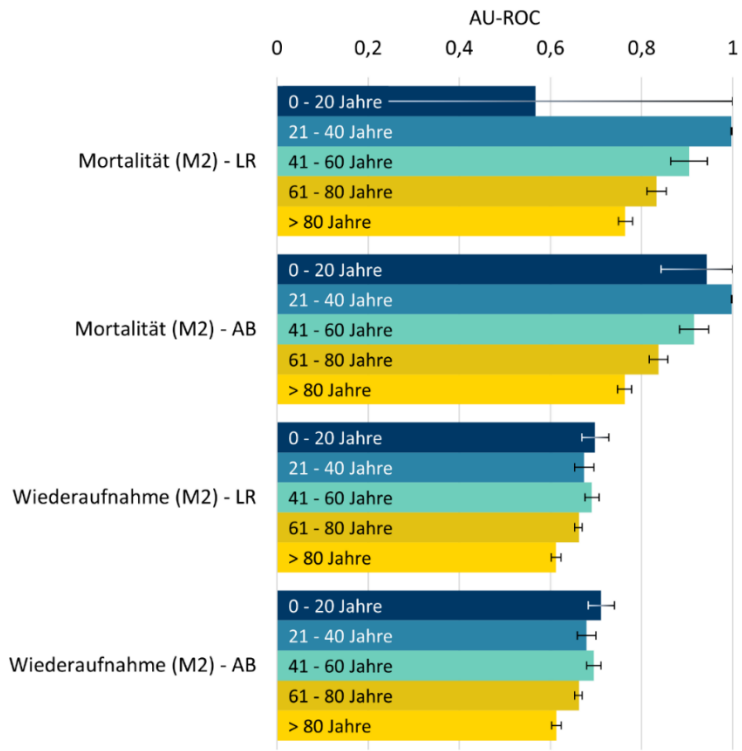

Abbildung 7: Prognosegüte (AUC-ROC) nach Altersgruppen für logistische Regression (LR) und AdaBoost (AB), Outcomes Mortalität und Ungeplante Wiederaufnahme

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Um die Güte der Subgruppen zu vergleichen, wurde die Fläche unter der ROC-Kurve (AUC-ROC) je Subgruppe berechnet. Exemplarisch werden die Ergebnisse für die Altersgruppen in Abbildung 7 dargestellt. In Altersgruppen zeigten sich insbesondere im Hinblick auf das Outcome Mortalität deutliche Unterschiede, mit niedrigeren AUC-ROC-Werten in den höheren Altersgruppen. Die Ergebnisse zum Outcome Mortalität in den beiden Altersgruppen bis 40 Jahre sind aufgrund einer geringen Fallzahl (nur 1 bis 2 Todesfälle je Gruppe) inhaltlich nicht interpretierbar und sollten daher außer Acht gelassen werden. Zudem können die verwendeten Teststatistiken in diesen Gruppen zu inadäquaten Vertrauensbereichen führen.

Weitere Subgruppenanalysen, beispielsweise nach Geschlecht, Pflegeheimbewohnerstatus oder Art des Krankenhausaufenthalts (Regel- vs. Notfallaufnahme) sind im Weißbuch (siehe Anlage 1, Kapitel 8.3.2) aufgeführt.

5 Diskussion der Projektergebnisse

5.1 Nutzbarkeit von GKV-Routinedaten für ML-Verfahren (Projektziel 1)

Im Laufe des Projektes konnten zahlreiche Erfahrungen gesammelt werden, inwiefern sich GKV-Routinedaten für den Einsatz von Verfahren des Maschinellen Lernens eignen. Zu den wesentlichen **Vorteilen von GKV-Routinedaten** ist zu zählen, dass sie aufgrund der gesetzlich vorgeschriebenen Erfassung und Verarbeitung in strukturierter und standardisierter Form vorliegen. Dies gilt nicht nur für Versicherte von verschiedenen Betriebskrankenkassen (wie im Projekt gezeigt), sondern grundsätzlich für alle gesetzlichen Krankenkassen in Deutschland. Darüber hinaus sind die Daten längsschnittlich verknüpfbar, so dass auch Analysen von mehrjährigen Beobachtungszeiträumen möglich sind. Zudem können Abrechnungsdaten aus verschiedenen SGB-Kontexten, beispielsweise aus dem ambulanten oder stationären Bereich, sektorenübergreifend verwendet werden. Positiv ist auch, dass insbesondere die abrechnungsessenziellen Daten weitgehend vollständig und in geprüfter Form bei den Krankenkassen vorliegen. Im analytischen bzw. ML-Kontext ist dies besonders vorteilhaft, weil die Daten nicht nur eine hohe Validität, sondern auch einen geringen Anteil an fehlenden Werten (engl. missings) aufweisen, welche ansonsten erst über aufwändige Verfahren (z. B. mit Imputationstechniken) korrigiert oder angepasst werden müssten. Auch im Projekt haben sich der große Datenumfang zu über einer Million Versicherten und die hohe Datenqualität als vorteilhaft erwiesen, um möglichst präzise Vorhersagemodelle entwickeln zu können.

Gleichzeitig konnten im Projekt auch **Nachteile von GKV-Routinedaten** bei der Anwendung von Methoden des Maschinellen Lernens festgestellt werden. So liegen die Abrechnungsdaten aus den verschiedenen SGB-Leistungsbereichen in einer relationalen Datenbank vor. Die Datenaufbereitung zu einem analysefähigen Datensatz gestaltete sich im Projekt sehr aufwändig und setzt nicht nur methodische Erfahrungen im Umgang mit Routinedaten, sondern auch Kenntnisse zur GKV-Abrechnungssystematik und zu Spezifika der deutschen Gesundheitsversorgung voraus. Zudem benötigten die Datenaufbereitung und die Umsetzung einzelner Ideen zur Optimierung der ML-Verfahren teilweise sehr viele personelle Ressourcen und Zeit. Dies erschwerte das explorative und nicht theoriegeleitete Analysieren der Daten, wie es im Bereich der Künstlichen Intelligenz verbreitet ist. Zudem enthielten die Analysedatensätze selbst bei einem theoriegeleiteten Vorgehen schnell eine Vielzahl potenziell relevanter Prädiktoren (vgl. Datenmodell M3 mit quartalsbezogenen Diagnosen),

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

weshalb einige ML-Methoden (z. B. bei der Hyperparameteroptimierung) mit der zur Verfügung stehenden Hardware eine Rechenzeit von bis zu 14 Tagen benötigten. Zudem erwiesen sich die ML-Verfahren im Vergleich zur logistischen Regression weniger robust gegenüber unbalancierten Daten, wie es beispielsweise beim selten auftretenden Outcome Mortalität der Fall ist. Hier mussten die Daten bzw. die ML-Verfahren mit Hilfe spezieller Methoden (Up- und Downsampling, Verlustfunktion) angepasst werden, um brauchbare Vorhersagen zu erzielen.

Über die im Projekt durchgeführten Analysen hinaus, die sich auf das Modelltraining und die Modellevaluation auf Basis „historischer" Routinedaten beschränken, lassen sich aus den Projekterkenntnissen auch **Implikationen für die Verwendung von GKV-Routinedaten** bei KI-gestützten Vorhersagemodellen oder anderen KI-Anwendungsmöglichkeiten in der Regelversorgung ableiten. So ist die Nutzbarkeit von Abrechnungsdaten vor allem dadurch limitiert, dass sie erst nach einem gewissen Zeitraum in geprüfter Form bei den Krankenkassen vorliegen und abgerufen werden können. Dieser Datenverzug variiert zwischen den SGB-Leistungsbereichen und kann beispielsweise bei ambulanten Abrechnungsdaten bis zu neun Monaten dauern. Diese eingeschränkte Datenaktualität gilt es sowohl für die Modellentwicklungsphase als auch bei der der späteren Berechnung von Prognosen „in Echtzeit" zu berücksichtigen. Darüber hinaus gelten die Routinedaten von Krankenkassen gemäß SGB als Sozialdaten und unterliegen somit hohen Auflagen. Derzeit werden Routinedaten abseits des primären Abrechnungszweckes entweder von den Krankenkassenkassen selbst auf der Grundlage gesetzlicher Vorgaben (vgl. § 284 SGB V, siehe Kapitel 6), genutzt oder es sind behördliche Genehmigungen (z. B. für Forschungs- oder Planungszwecke) notwendig, deren Beantragung erfahrungsgemäß mehrere Monate dauert und die Arbeit – vor allem in Forschungsprojekten mit einer begrenzten Laufzeit – zuweilen erschwert. Zudem sieht die Gesetzgebung eine Einwilligung der Versicherten für die Verwendung ihrer Daten vor, was sich gerade bei den Versichertengruppen, die besonders von KI-Lösungen profitieren könnten (z. B. multimorbide oder pflegebedürftige Menschen), als Hürde erweist. Hier bleibt abzuwarten, ob und inwieweit das Gesundheitsdatennutzungsgesetz (GDNG) künftig die Nutzbarkeit der GKV-Routinedaten für Verfahren der Künstlichen Intelligenz verbessert.

5.2 Evaluation der Regressions- und ML-Verfahren (Projektziel 2)

5.2.1 Vergleich der Modellgüte

Bei der Betrachtung der Ergebnisse zur Modellgüte lohnt sich zunächst ein Blick auf die **Unterschiede bei den verschiedenen Modellvarianten (M1 bis M3)**, anhand derer die Regressions- und ML-Verfahren trainiert worden sind. So wurden drei unterschiedlich umfangreiche Sets an Variablen bzw. Features verwendet, um zu untersuchen, wie sich die Vorhersagen verändern, wenn verschiedene Modelle auf Basis unterschiedlich umfangreicher Informationen trainiert werden. Aus dem Vergleich der drei Modellvarianten geht hervor, dass bereits das Modell 1 (Basisprädiktoren) eine relativ hohe Erklärungskraft für das Eintreten der beiden Outcomes aufweist. Mit dem Modell 2 konnte die Vorhersagegenauigkeit erwartungsgemäß weiter verbessert werden, indem neben den Basisprädiktoren zusätzlich Vorerkrankungsdiagnosen berücksichtigt wurden. Für Modell 3 wurden die Informationen zu den Vorerkrankungen zusätzlich um die Quartale ihrer Dokumentation (vor der Krankenhausbehandlung) erweitert, womit sich die Vorhersagegüte der Modelle allerdings

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

nicht verbessern ließ, sondern zum Teil sogar verschlechterte. Das Modell 2 wurde vor diesem Hintergrund als das beste und somit finale Modell für die vergleichende Testung ausgewählt. An dieser Stelle lässt sich konstatieren, dass mehr Informationen im Rahmen von ML-Analysen nicht zwangsläufig zu besseren Vorhersageergebnissen führen.

Von zentraler Bedeutung sind in dem Projekt die **Ergebnisse zur Modellevaluation der verschiedenen ML-Verfahren** im Vergleich zur klassischen logistischen Regression. Die Testung der eingesetzten Vorhersageverfahren auf Basis identischer Daten zeigt, dass das ML-Verfahren AdaBoost und die logistische Regression die höchste Genauigkeit bei der Vorhersage der Mortalität und der Ungeplanten Wiederaufnahme erzielen. Die Unterschiede (zugunsten AdaBoost) sind allerdings eher gering und dürften für die praktische Umsetzung zunächst kaum Relevanz besitzen. Im Vergleich dazu schneiden Random Forest und Neuronale Netze in der hier gewählten Implementierung identischer Daten (gemäß den drei Modellvarianten M1 bis M3) bei beiden Outcomes schlechter ab. Hierzu ist anzumerken, dass das sogenannte „Feature Engineering“, also die Umwandlung und Anreicherung von Daten mit speziellen Verfahren, eine zentrale Rolle für die Leistungsfähigkeit von ML-gestützten Prädiktionsmodellen spielt. Das Potenzial komplexerer Verfahren, wie von Neuronalen Netzen, wurde so möglicherweise noch nicht vollständig ausgeschöpft. Zukünftige Analysen sollten daher auch auf erweitertes Feature Engineering setzen. Dabei könnte auch bei der Verwendung von GKV-Routinedaten ein großes Potenzial in der Anwendung aktueller, auf Transformer-Modellen basierender Daten-Enkodierungen (z. B. zur Verdichtung von ICD-Diagnosen) liegen, da diese in der Lage sind, komplexe Abhängigkeiten in den Daten zu erfassen. Zu diesem Thema ist im Rahmen des Projektes KI-THRUST eine Masterarbeit entstanden, in der ein Transformer-Modell (Bert) und ein Word2Vec-Modell (Pat2Vec) auf der Grundlage von ICD-Diagnosen angewendet und mit dem Datenmodell M1 verglichen wurden. Die Ergebnisse der Arbeit weisen darauf hin, dass vor allem das Bert-Modell in der Lage ist, bessere Vorhersagen als die anderen Modelle zu erzielen (siehe Kapitel 7, Harriehausen et., 2025; geplante Veröffentlichungen: Harriehausen et al.).

5.2.2 Unterschiede in der Vorhersagbarkeit der Outcomes

Die im Projekt exemplarisch betrachteten Outcomes Mortalität und Ungeplante Wiederaufnahme lassen sich mit den eingesetzten ML-Verfahren unterschiedlich gut vorhersagen. Anhand der verwendeten Evaluationsmetrik zur Fläche unter der ROC-Kurve (AUC-ROC) lässt sich für das Outcome Mortalität feststellen, dass alle Modelle formal eine „ausgezeichnete“ Vorhersagegüte (AUC-ROC > 0,8) erreichen. Im Gegensatz dazu liegt die Vorhersagequalität für das Outcome Ungeplante Wiederaufnahme bei allen Verfahren formal unter einer gemeinhin als „akzeptabel“ angenommenen Schwelle (AUC-ROC < 0,7). Das Outcome Mortalität ist demnach deutlich besser anhand der hier genutzten Routinedaten der Krankenkassen vorhersagbar als das Outcome Ungeplante Wiederaufnahme. Weitere Hinweise zur Praktikabilität liefern die im Projekt ebenfalls berechneten Flächenwerte unter der Precision-Recall-Kurve (AUC-PR), da sie die erwartbare Häufigkeit eines Outcomes in Populationen berücksichtigen und somit Abschätzungen zu Konsequenzen bzw. zum Einsatz von Interventionen, die für bestimmte Vorhersagen geplant sind, erlauben. Entsprechende Ergebnisse zeigen, dass die Präzision der Modelle für beide Outcomes eher gering ausfällt, also auch Personen aus „Risikogruppen“ mit vergleichsweise hohen vorhergesagten Risiken nachfolgend häufig real nicht von den jeweils betrachteten Outcomes betroffen sind. Ein

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

wesentlicher Grund hierfür liegt in der starken Unbalanciertheit des Datensatzes bzw. der niedrigen Prävalenz der positiven Klassen, die auch in vielen epidemiologischen Studien dazu führt, dass sich zwar Risikofaktoren bzw. Merkmalsausprägungen identifizieren lassen, die mit teils stark erhöhten relativen Risiken assoziiert sind, aber auch Personen mit diesen Risikofaktoren innerhalb begrenzter Zeiträume eher selten betroffen sind. Ein plakatives Beispiel hierfür ist, dass auch Kettenraucher mehrheitlich die kommenden 365 Tage überleben. Daraus lässt sich schlussfolgern, dass die im Projekt entwickelten Prognosemodelle zwar in der Lage sind, komplexe Zusammenhänge in umfangreichen Daten zu erfassen und ausreichend präzise Vorhersagen auf der Populationsebene zu erzielen. In Anbetracht der genannten Einschränkungen sollten die Modelle jedoch nur mit Bedacht in der Praxis für die individuelle Prognostik verwendet werden und nicht die alleinige Entscheidungsgrundlage (z. B. für therapeutische Maßnahmen) bilden.

5.2.3 Übertragbarkeit und Fehlklassifikationen

Im Projekt wurde zusätzlich untersucht, inwieweit das Regressionsmodell und das AdaBoost-Modell (als bestes ML-Verfahren), die jeweils auf Daten aus dem Jahr 2018 entwickelt worden sind, für Vorhersagen in Daten aus den Folgejahren 2019 und 2020 genutzt werden können. Die Nutzbarkeit in Folgejahren ist insofern besonders wesentlich, als dass sie dem typischen Anwendungsfall entspricht, in dem ein mit verfügbaren Daten und bekannten Outcomes entwickeltes Modell mit neu erfassten Daten genutzt wird (bei dem in der Praxis noch niemand das Outcome kennt). Gemessen an AUC-ROC-Werten erwiesen sich beide Modelle auch für das Jahr 2019 als geeignet, tendenziell lagen hier die Werte der ermittelten Gütemaße sogar noch etwas höher. Bei einer Anwendung der Vorhersagen für Behandlungsfälle im Jahr 2020 zeigten sich demgegenüber reduzierte Gütemaße, was aufgrund der gravierenden Auswirkungen der COVID-19-Pandemie (v. a. auf die stationäre Versorgung) den Erwartungen entspricht. Die Veränderung der Gütemaße bewegte sich jedoch auch hier nur in einem begrenzten Rahmen, womit die Nutzbarkeit der Vorhersagemodelle auch im ersten Pandemiejahr allenfalls graduell eingeschränkt war. Dieses Ergebnis übertrifft anfängliche Erwartungen im positiven Sinne, da es in den Jahrzehnten vor 2020 in Deutschland kaum ähnlich gravierende Einschnitte bezüglich der gesundheitlichen Versorgung gegeben haben dürfte und insofern auch deutliche Einschränkungen bei der Prognosegüte hätten erwartet werden können.

Darüber hinaus wurden im Projekt verschiedene Subgruppenanalysen durchgeführt, um innerhalb der stationär behandelten Versicherten Personengruppen zu identifizieren, die einen überdurchschnittlich hohen Anteil an fehlklassifizierten Fällen aufweisen. Gemessen an AUC-ROC-Werten zeigten sich eingeschränkte Prognosemöglichkeiten insbesondere innerhalb der Gruppen von Personen im Alter von über 80 Jahren sowie bei Personen mit längerfristiger Unterbringung in Pflegeheimen. Die „schlechteren“ Prognosemöglichkeiten innerhalb dieser Gruppen dürften maßgeblich daraus resultieren, dass innerhalb von entsprechenden „Hochrisikogruppen“ naturgemäß die Varianz relevanter Risikofaktoren begrenzt ist und verbleibende Risikofaktoren die Risiken innerhalb dieser Gruppen nur noch eingeschränkt differenzieren können. Wesentlich erscheint die Beobachtung, dass Veränderungen der Vorhersagegüte in Subgruppen bei beiden Verfahren (logistische Regression und AdaBoost) sehr ähnlich sind, womit sich keine Hinweise auf Abnormitäten der Vorhersagen in Subgruppen bei einem der beiden Modelle ergeben.

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

5.2.4 Erklärbarkeit

Im Gegensatz zum logistischen Regressionsmodell, dessen Vorhersageergebnisse sich anhand von Regressionskoeffizienten und anderer Kennzahlen gut interpretieren lassen, sind komplexere maschinelle Lernverfahren oft weniger transparent in Bezug auf die Bedeutung einzelner Prädiktoren (z. B. Versichertenmerkmale, Diagnosen). Den Beitrag eines Merkmals zur Vorherhersage lässt sich bei solchen Modellen jedoch mithilfe erklärender Verfahren, wie Shapley Value Sampling, quantifizieren. Im Projekt wurden für die berechneten ML-Verfahren verschiedene Methoden zur Erklärbarkeit getestet, um die Bedeutung einzelner Prädiktoren bewerten zu können. Die entsprechenden Erklärbarkeitsanalysen haben gezeigt, dass Integrated Gradients im Vergleich zu Shapley Value Sampling und LIME weniger zuverlässige Relevanzzuweisungen liefert. Dies liegt vermutlich an seiner starken Abhängigkeit vom Gradienten und dem Modelloutput, die bei kleinen Werten zu kaum interpretierbaren Ergebnissen führen. Im Gegensatz dazu identifizieren LIME und Shapley konsistent relevante Merkmale, insbesondere schwerwiegende Vorerkrankungen, die inhaltlich plausibel mit den Outcomes Ungeplante Wiederaufnahme und Mortalität zusammenhängen. Trotz dieser ersten Einblicke bleibt die Erklärbarkeit der Modellvorhersagen begrenzt, sodass derzeit keine verlässlichen Aussagen darüber getroffen werden können, warum eine spezifische Patientin oder ein Patient eine bestimmte Vorhersage erhält.

5.2.5 Implementierbarkeit

Im Projekt wurden grundlegende Überlegungen angestellt, auf welche Weise die erprobten Vorhersageverfahren technisch implementiert werden könnten, um beispielsweise als Entscheidungsunterstützungssysteme in Krankenhäusern genutzt werden zu können. Die Implementierung von Regressionsmodellen zur Unterstützung des Krankenhaus-Entlassmanagements konnte bereits erfolgreich im Vorläuferprojekt USER demonstriert werden (Broge et al., 2024). In diesem Projekt wurden automatisierte Prozeduren in die IT-Infrastruktur der Krankenkassen implementiert, um aus den Routinedaten die für die Modelle benötigten Eingabedaten (konkret die Prädiktorwerte) in Echtzeit zu verarbeiten. Anschließend wurden diese Informationen in mathematische Regressionsgleichungen, die den Regressionsmodellen zugrunde liegen, übertragen und für die Berechnung der individuellen Vorhersagewahrscheinlichkeiten genutzt. Sämtliche Verarbeitungsprozesse erfolgten innerhalb der Krankenkasseninfrastruktur, so dass keine personenbezogenen Abrechnungsdaten, sondern nur die Prognosescores mit Zustimmung der Versicherten an die Krankenhäuser übermittelt wurden. Bei einer Implementierung der ML-Verfahren könnte die automatisierte Datenverarbeitung zunächst analog zum Projekt USER erfolgen. Wesentliche Unterschiede ergeben sich allerdings bei der Berechnung der Vorhersagen. Hierfür ist ein spezieller Softwarebereitstellungsprozess (sog. Containerisierung) erforderlich, bei dem ein Code bzw. Algorithmus eines trainierten Modells mit allen dazugehörigen Dateien und Bibliotheken gebündelt wird, die für die Ausführung in einer beliebigen Infrastruktur benötigt werden. Über eine entsprechende Anwendungsschnittstelle (API) könnten dann externe Systeme (z. B. der Krankenkassen) Eingabedaten einspeisen und die ermittelten Vorhersagen abrufen. Im Gegensatz zum Modelltraining, wären die dabei anfallenden Leistungsanforderungen und Rechenzeiten bei den ML-Verfahren wahrscheinlich nicht bedeutsam höher als bei den Regressionsmodellen. Anschließend wäre es möglich, die Vorhersageergebnisse entweder an das Krankenhausinformationssystem (KIS) oder an eine

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

separate Softwareplattform mit Schnittstelle zum KIS zu übermitteln und im Krankenhaus über eine entsprechende Benutzeroberfläche abzurufen. In dieser Oberfläche sollten die Vorhersageergebnisse möglichst anwenderfreundlich (im Sinne von übersichtlich, verständlich und selbsterklärend) dargestellt werden. Da reine Eintrittswahrscheinlichkeiten (zumeist in Prozent angegeben) häufig schwierig für Anwenderinnen und Anwender zu interpretieren sind, gilt es zu überlegen, ob zusätzliche Interpretationshilfen sinnvoll sind. Beispiele hierfür sind die Angabe der prognostizierten Klasse (z. B. Wiederaufnahme in 30 Tagen: ja/nein), Warnhinweise (sog. Red Flags) oder Ampelanzeigen. Für diese Darstellungen werden allerdings zugrundeliegende Schwellenwerte benötigt, deren Festlegung in der Regel mit ökonomischen, juristischen und ethischen Implikationen einhergeht.

5.3 Weißbuch (Projektziel 3)

Das fertige Weißbuch (siehe Anlage 1) wurde als Online-Publikation vom aQua-Institut veröffentlicht und ist über folgende Webadressen abrufbar:

- Permalink: https://www.aqua-institut.de/fileadmin/aqua_de/PermaLink/KI-THRUST-Weissbuch-KI-und-GKV-Routinedaten.pdf.
- Persistent Identifier (PID): https://hdl.handle.net/21.11101/0000-0007-FFB6-D.

Die Webadressen wurden und werden auch künftig vom Projektkonsortium über verschiedene Kanäle (via Webseiten, Social-Media-Plattformen, Pressemitteilungen etc.) und Netzwerkaktivitäten (z.B. in Fachgesellschaften) verbreitet, um möglichst viele potenzielle Leserinnen und Leser, vor allem aus den Bereichen Gesundheit und KI, erreichen und auf das Weißbuch aufmerksam machen zu können. Korrekturen und Überarbeitungen sind auch über die Veröffentlichung und das Projektende hinaus angedacht.

6 Verwendung der Ergebnisse nach Ende der Förderung

Veröffentlichung des Weißbuchs: Technisch-methodische Orientierung für Akteure der Sekundärdatennutzung

Das im Projekt verfasste Weißbuch soll über die Projektlaufzeit hinaus als zentraler Zugangspunkt zur methodischen, technischen und organisatorischen Aufarbeitung von KI-Verfahren auf der Grundlage von Routinedaten der gesetzlichen Krankenversicherung fungieren. Die Veröffentlichung ist in digitaler Form erfolgt und richtet sich an Fachöffentlichkeiten, die an der Entwicklung, Evaluation oder Implementierung von maschinellen Lernverfahren unter Verwendung GKV-seitiger Datenbestände interessiert sind.

Inhaltlich bietet das Weißbuch eine praxisnahe Einführung in Methoden des Maschinellen Lernens, die speziell auf Routinedaten der Gesetzlichen Krankenversicherung (GKV) angewendet wurden. Besonderes Augenmerk liegt auf den Rahmenbedingungen für die Nutzung solcher Verfahren im Gesundheitswesen: Datenauswahl, Modellbildung, Validierung, aber auch organisatorische Überlegungen zur Implementierung in Versorgungsprozesse werden erörtert. Das Dokument geht dabei über eine rein technische Handreichung hinaus. Es enthält Abschnitte zu Fragen der nötigen Ressourcen, der rechtlichen Voraussetzungen und der unternehmensstrategischen Integration von KI-Projekten. Alle für den Erfolg des Projekts wesentlichen Schritte wurden so transparent wie möglich dargestellt. Dabei wird auf so

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

unterschiedliche Themen, wie die Auswahl von Softwarebibliotheken oder die Beantragung der Übermittlung von Sozialdaten für die Forschung gem. § 75 SGB X, eingegangen und in einen gemeinsamen und nachvollziehbaren Kontext gesetzt. Das Konsortium hatte dabei das Zielbild, die Schritte reproduzierbar und das Projekt insgesamt für künftige Forschungsvorhaben bzw. Folgeprojekte adaptierfähig zu machen.

Nutzen für weiterführende Forschung und Folgeprojekte

Das Projekt KI-THRUST war auf methodische Grundlagenentwicklung ausgerichtet. Die zentrale Zielsetzung bestand darin, exemplarisch zu überprüfen, ob sich GKV-Routinedaten für Verfahren des Maschinellen Lernens eignen und ob sich unter Berücksichtigung regulatorischer Anforderungen ein verallgemeinerbarer methodischer Rahmen entwickeln lässt. Das Projekt ist damit primär explorativ und erkenntnisorientiert ausgerichtet.

Die Erkenntnisse lassen sich primär als konzeptionelle Vorarbeiten für weiterführende empirische Forschungsvorhaben einstufen. Diese könnten etwa die Übertragbarkeit der Modelle nicht nur auf das Entlassmanagement, sondern auch auf andere Prozesse und Anwendungssettings vereinfachen. Ferner könnte das methodische Vorgehen im Projekt Grundlage dafür sein, andere Outcomes vorhersagen zu können, wie Reha- oder Pflegebedarfe oder die Entstehung bestimmter (vermeidbarer) Erkrankungen.

Besonders in den Fällen, in denen klassische Verfahren, wie die logistische Regression, an methodische Grenzen stoßen oder nicht ausreichend erklärbare Ergebnisse liefern, sollten alternative Modellierungsansätze mithilfe der Erkenntnisse aus KI-THRUST erprobt werden.

Auch auf konzeptioneller Ebene gibt es Forschungsbedarf: Neue Input-Output-Theorien könnten zur Entwicklung besserer Zielgrößen beitragen, da viele Versorgungsprozesse nicht linearen, sondern verschachtelten Pfaden mit vielen Zwischenstationen und Rückkopplungseffekten folgen, die sich in traditionellen Modellstrukturen nur schwer abbilden lassen. Die Vorarbeiten aus KI-THRUST – Aufbereitung von Routinedaten und das Wissen über KI-Verfahren und Datennutzungsprozesse – könnten hierfür eine deutliche „Abkürzung“ bedeuten.

Zumindest das AdaBoost-Verfahren zeigte eine leichte Verbesserung gegenüber der logistischen Regression. Das Kosten-Nutzen-Verhältnis spricht aktuell aber noch zugunsten des klassischen Verfahrens. Solange KI-Verfahren noch vergleichsweise teuer und aufwändig sind, dürfte der Vorteil geringer als die zusätzlichen Kosten ausfallen. Mit zunehmendem technischem Fortschritt könnte der geringe Vorteil jedoch in wenigen Jahren entscheidungsrelevant sein. Deshalb sollten Krankenkassen sich bereits heute ausreichend Wissen aneignen und Ressourcen für diesen Bereich ausbauen. Dafür stellt das Projekt KI-THRUST und das Weißbuch eine erste Orientierungshilfe dar.

Anwendungsfälle für KI-Verfahren in der Versorgungskonzeption

Die Anwendung von KI-basierten Prognosemodellen auf Basis von GKV-Routinedaten ist grundsätzlich für eine Vielzahl von versorgungsnahen Fragestellungen denkbar. Die methodische Eignung ergibt sich insbesondere für Kontexte, in denen retrospektiv strukturierte, annotierte Daten mit bekannter Zielvariablenbeziehung vorliegen. Dies ist typischerweise bei Leistungsdaten aus Krankenhausaufenthalten, Rehabilitationsmaßnahmen oder Arzneimittelverordnungen der Fall.

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Ein geeigneter Anwendungsfall besteht – ungeachtet der Berechnungsergebnisse in diesem Projekt – weiterhin im Bereich des Entlassmanagements nach § 39 Abs. 1a SGB V. Die sektorenübergreifende Organisation nachstationärer Versorgungsbedarfe erfordert eine frühzeitige Einschätzung potenzieller Folgebedarfe wie Rehabilitation, Pflege, Hilfsmittelversorgung oder medizinische Anschlussleistungen. Prognosemodelle können hierzu ergänzende Entscheidungsgrundlagen liefern, indem sie auf Basis strukturierter Falldaten Wahrscheinlichkeiten für spezifische Folgebedarfe berechnen. Dass dies in der Praxis gut funktionieren kann, wurde im Vorläuferprojekt USER gezeigt. Es ist problemlos vorstellbar, die in jenem Projekt verwendete Schnittstelle zwischen den Krankenhäusern und Krankenkassen und den darüber übermittelten individualisierten Versorgungsbedarfsscore bei der Aufnahme von Patientinnen und Patienten auch mit KI-Verfahren zu berechnen und dem behandelnden Krankenhauspersonal zur Verfügung zu stellen.

Solche Anwendungsfälle lassen sich konzeptionell auf andere Versorgungsbereiche übertragen, sofern geeignete Zielvariablen, ein stabiler Datenzugang sowie belastbare Theorien zur Input-Output-Beziehung bestehen. Dazu zählen z. B. Fragestellungen der Risikostratifizierung bei chronischen Erkrankungen, die Versorgungspfade multimorbider Patientengruppen oder indikationsspezifische Kombinationsmuster von Arznei- und Hilfsmittelversorgung. Insbesondere bei Versorgungsprogrammen mit strukturierter Dokumentation (Disease-Management-Programme, sektorenübergreifende Versorgungskonzepte) besteht potenziell eine hohe Anschlussfähigkeit für die prädiktive Modellierung.

Nutzung von Prädiktion entlang der Patient Journey

Die Anwendung prädiktiver Verfahren entlang der gesamten Patient Journey bietet zahlreiche Ansätze für individuell angepasste Verläufe durch präzisere Vorhersagen und damit einhergehend Potenziale zur Effizienzsteigerung und Kostendämpfung in der Gesundheitsversorgung. Grundlage sind jeweils datengestützte Risikostratifizierungen, die sowohl zur Identifikation relevanter Zielgruppen als auch zur Steuerung konkreter Interventionspfade dienen können. Als rechtliche Grundlagen sind insbesondere die neueren §§ 25b und 68b SGB V zu diskutieren, die unter bestimmten Voraussetzungen datengestützte Analysen und die direkte Ansprache der Versicherten erlauben.

Im Rahmen des § 25b SGB V können etwa Versicherte identifiziert werden, bei denen ein erhöhtes Risiko für schwerwiegende Erkrankungen oder Komplikationen besteht, beispielsweise bei Arzneimitteltherapien oder Pflegeverläufen. Prognosemodelle auf Basis Maschinellen Lernens, die erklärbar sein müssen, könnten hier genutzt werden, um Empfehlungen für präventive ärztliche Konsultationen zu generieren. Die technische Umsetzung könnte dann über ein Scoring-System mit definierter Schwellenwertlogik sowie geeigneten Mechanismen zur automatisierten Benachrichtigung der Versicherten erfolgen.

Der § 68b SGB V ermöglicht es Krankenkassen, datenbasierte Analysen für die Entwicklung neuer Versorgungsangebote durchzuführen und auf dieser Grundlage individuelle Angebote zu unterbreiten. KI-gestützte Verfahren könnten hier, soweit zulässig, genutzt werden, um diejenigen Versicherten zu identifizieren, für die ein hohes Maß an potenziellem Nutzen besteht.

Auch bei standardisierten Versorgungspfaden, z. B. bei Standard Operating Procedures (SOP) in Kliniken, besteht Anwendungspotenzial für prädiktive Modellierungen mit GKV-

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

Routinedaten. Insbesondere in der perioperativen Versorgung oder in Rehabilitationsverläufen könnten Risikoabschätzungen genutzt werden, um zeitliche Abläufe, Frequenzen von Kontrolluntersuchungen oder die Zuweisung zu Nachsorgeeinrichtungen individuell zu adaptieren. Prognosen könnten in dem Zusammenhang auch die Planung von Betten, Geräten, Personal und weiteren Ressourcen erleichtern. Voraussetzung hierfür ist eine strukturierte Dokumentation und eine regelhafte Verfügbarkeit der relevanten Prozessdaten.

Zusätzliche Anwendungsperspektiven ergeben sich durch die Kombination von GKV-Routinedaten mit klinischen Daten aus der elektronischen Patientenakte (ePA) oder dem Forschungsdatenzentrum Gesundheit. Die Integration beider Datenquellen erlaubt die Entwicklung umfassenderer Patientenmodelle mit höherer Vorhersagegüte und kann zur Entwicklung adaptiver Versorgungspfade beitragen. Hier sollte KI-THRUST ebenfalls bereits einen Beitrag zu den Vorarbeiten darstellen.

7 Erfolgte bzw. geplante Veröffentlichungen

Originalarbeiten/ Veröffentlichungen

- Grobe, T., Pollmann, T., Ramcke, D., Weller, L., Kretzler, M., Hauschild, A.-C., Maurer, M. C., Metsch, J. M., Ritter, Z. (2025). Weißbuch – Potenziale KI-gestützter Vorhersageverfahren auf Basis von GKV-Routinedaten. Göttingen: aQua-Institut. https://www.aqua-institut.de/fileadmin/aqua_de/PermaLink/KI-THRUST-Weissbuch-KI-und-GKV-Routinedaten.pdf.

Beiträge für wissenschaftliche Veranstaltungen (chronologisch sortiert)

- Pollmann, T.: Prädiktionsmodelle aus Routinedaten der Krankenkassen. DMEA-Satellitenveranstaltung 2023 von GMDS und BVMI. Berlin, 24.04.2023.
- Pollmann, T., Hauschild, A-C.: KI-gestützte Vorhersageverfahren auf Basis von GKV-Routinedaten - Erkenntnisse aus dem Projekt KI-THRUST. 26. Arbeitstreffen der User Group „Analytik in der Krankenversicherung“ der Gesundheitsforen Leipzig. Leipzig, 12.10.2023.
- Weller, L., Pollmann, T., Starke, P., Grobe, T.: KI-gestützte Vorhersageverfahren auf der Basis von GKV-Routinedaten – Erste Ergebnisse aus dem Projekt KI-THRUST. AGENS-Methodenworkshop 2024. Hannover, 20.03.2024.
- Weller, L., Pollmann, T., Starke, P., Grobe, T.: KI-gestützte Vorhersageverfahren mit strukturierten GKV-Routinedaten - Erfahrungen aus der Praxis und Perspektiven. Veranstaltung „KI und Routinedaten“ der AGENS und AG Validierung und Linkage von Sekundärdaten (DNVF). Online, 03.06.2024.
- Pollmann, T., Weller, L., Starke, P., Metsch, JM., Maurer, MC., Ritter, Z., Hauschild, A-C., Kretzler, M., Grobe, T.: Alles besser mit KI? Ein Vergleich von ML-Methoden und klassischen Regressionsmodellen zur Vorhersage poststationärer Ereignisse. 23. Deutscher Kongress für Versorgungsforschung. Potsdam, 25.09.2024. doi: 10.3205/24dkvf218

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

- Kretzler, M.: KI-THRUST - Potenziale KI-gestützter Vorhersageverfahren auf Basis von Routinedaten (Vortrag). GKV-Expertise beim GKV-Spitzenverband. Berlin, 21.11.2024
- Harriehausen, J., Metsch, JM., Ritter, Z., Maurer, MC., Weller, L., Pollmann, T., Kretzler, M., Grobe, T., Hauschild, A-C.: Evaluating Transformer Models for ICD Code Embeddings in Predicting Clinical Outcomes. 3rd Hiedelberg Spring Symposium Medical Informatics. Heidelberg, 28.05.2025
- Ritter, Z., Maurer, MC., Metsch, JM., Weller, L., Pollmann, T., Kretzler, M., Grobe, T., Hauschild, A-C.: Potential of Machine Learning for Discharge Management Using Routine Health Insurance Data. 70. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS). Jena, 7-11.09.2025 (akzeptiert, Poster)

Öffentlichkeitsarbeit

- Projektvorstellung auf Webseiten
  - https://www.aqua-institut.de/projekte/ki-thrust
  - https://zdin.de/kooperationsmoeglichkeiten/ki-thrust-praediktion-von-ereignissen-bei-routinedaten
- Kretzler, M. (2024): Schubkraft durch KI-Trainingsdaten. In: Betriebskrankenkassen Magazin des BKK Dachverbands (BKK-Magazin), 06/2024, S. 24-31., https://www.bkk-dachverband.de/fileadmin/Artikelsystem/Magazin/2024/Heft_6/BKK_06_2024_gesamt_web.pdf, abgerufen am 29.07.2025
- Abschlusssymposium „BKK Innovativ: KI THRUST – Bedarfe Erkennen und Gesundheit gestalten mit Methoden der künstlichen Intelligenz“, am 20.11.2024 (Online), Aufzeichnung verfügbar unter: https://www.bkk-dachverband.de/innovation/bkk-innovativ/ki-thrust, abgerufen am 29.07.2025.

Geplante Veröffentlichungen

- Harriehausen et al. (in Vorbereitung). Transformer Encodings in BioMedical Data (Arbeitstitel). Zur Einreichung bei einem peer-reviewten Journal im Bereich Medical Informatics vorgesehen.

IV Literaturverzeichnis

Broge, B., Kleine-Budde, K., Pollmann, T., Blum, K., Finger, B. (2020). Ergebnisbericht EMSE – Entwicklung von Methoden zur Nutzung von Routinedaten für ein sektorenübergreifendes Entlassmanagement. https://innovationsfonds.g-ba.de/downloads/beschluss-dokumente/8/2020-04-03_EMSE_Ergebnisbericht.pdf.

Broge, B., Lingnau, R., Pollmann, T., Willms, G., Blum, K. (2024). Ergebnisbericht USER – Umsetzung eines strukturierten Entlassmanagements mit Routinedaten. https://innovationsfonds.g-ba.de/downloads/beschluss-dokumente/554/2024-04-19_USER_Ergebnisbericht.pdf.

Gefördert durch: Gemeinsamer Bundesausschuss Innovationsausschuss

V Anlagen

Anlage 1:	Weißbuch (Titel: „Weißbuch - Potenziale KI-gestützter Vorhersageverfahren auf Basis von GKV-Routinedaten“), Stand 07/2025
Anlage 2:	Konzept für eine projektspezifische IT-Infrastruktur, Stand 12/2021
Anlage 3:	Datensatzbeschreibung zum Projekt KI-THRUST, Stand 23.05.2022

Gefördert durch:
Gemeinsamer Bundesausschuss Innovationsausschuss

Anlage 1: Weißbuch (Titel: „Weißbuch - Potenziale KI-gestützter Vorhersageverfahren auf Basis von GKV-Routinedaten“), Stand 07/2025

Weißbuch

Potenziale KI-gestützter Vorhersageverfahren auf Basis von GKV-Routinedaten

Thomas G. Grobe, Anne-Christin Hauschild, Matthias Kretzler, Miriam C. Maurer, Jacqueline M. Metsch, Thorsten Pollmann, David Ramcke, Zully Ritter, Lisa Weller

Impressum

Weißbuch – Potenziale KI-gestützter Vorhersageverfahren auf Basis von GKV-Routinedaten

Autorenschaft (nach Institutionen und in alphabetischer Reihenfolge):
Thomas G. Grobe, Thorsten Pollmann, David Ramcke, Lisa Weller (jeweils aQua-Institut), Matthias Kretzler (BKK Dachverband), Anne-Christin Hauschild, Miriam Cindy Maurer, Jacqueline Michelle Metsch, Zully Ritter (jeweils Universitätsmedizin Göttingen)

Herausgeber: aQua – Institut für angewandte Qualitätsförderung und Forschung im Gesundheitswesen GmbH

Stand: Juli 2025

DOI: in Beantragung

Zitierhinweis: Grobe, T., Pollmann, T., Ramcke, D., Weller, L., Kretzler, M., Hauschild, A.-C., Maurer, M., Metsch, J., Ritter, Z. (2025). *Weißbuch – Potenziale KI-gestützter Vorhersageverfahren auf Basis von GKV-Routinedaten*. Göttingen: aQua-Institut.

Förderung: Dieses Weißbuch ist im Rahmen des Projektes KI-THRUST entstanden, welches mit Mitteln des Innovationsausschusses beim Gemeinsamen Bundesausschuss unter dem Förderkennzeichen 01VSF20014 gefördert wurde.

Danksagung: Das Projektkonsortium von KI-THRUST bedankt sich bei den projektteilnehmenden Krankenkassen, namentlich bei der BAHN-BKK, Novitas BKK, Pronova BKK und Siemens-Betriebskrankenkasse, für die Unterstützung des Projektes.

Die Verfasser haben große Mühe darauf verwandt, die fachlichen Inhalte auf den Stand der Wissenschaft bei Drucklegung zu bringen. Dennoch sind Irrtümer oder Druckfehler nie auszuschließen.

Für die Inhalte externer Links und fremder Webseiten übernehmen wir keine Haftung, da wir keinen Einfluss auf deren Inhalte haben. Für die Inhalte der verlinkten Seiten sind ausschließlich deren Betreiber verantwortlich.

Wir haben uns bemüht, eine gendergerechte Sprache umzusetzen und gleichzeitig die Lesbarkeit und Barrierefreiheit der Texte zu gewährleisten. Im Text werden daher möglichst neutrale Formen verwendet. Ist dies nicht möglich, werden an manchen Stellen (z. B. bei direkter Ansprache) die weibliche und männliche Form ausgeschrieben. Darüber hinaus wird weiterhin das generische Maskulinum verwendet. Personenbezeichnungen beziehen sich dann auf alle Geschlechter.

Inhalt

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AB	AdaBoost
AI	Artificial Intelligence (Künstliche Intelligenz)
ATC	Anatomisch-Therapeutisch-Chemisches Klassifikationssystem
BAS	Bundesamt für Soziale Sicherung
BSNR	Betriebsstättennummer
DDD	Daily Defined Dose
DRG	Diagnosis Related Groups
DVG	Digitale-Versorgung-Gesetz
EBM	Einheitlicher Bewertungsmaßstab
GDNG	Gesundheitsdatennutzungsgesetz
GDPR	General Data Protection Regulation (Datenschutz-Grundverordnung der EU)
GKV	Gesetzliche Krankenversicherung
HIPAA	Health Insurance Portability and Accountability Act
ICD	International Classification of Diseases
iGeL	Individuelle Gesundheitsleistungen
IK	Institutionskennzeichen
KI	Künstliche Intelligenz
KNN	Künstliches Neuronales Netz
KV	Kassenärztliche Vereinigung
LANR	Lebenslange Arztnummer
LR	Logistische Regression
MCAR	Missing Completely At Random
MCC	Matthews-Korrelationskoeffizient
ML	Machine Learning (engl.) / Maschinelles Lernen (dt.)
MLP	Multilayer Perceptron
OPS	Operationen- und Prozedurenschlüssel
OR	Odds Ratio
PEPP	Pauschalisierendes Entgeltsystem für Psychiatrie und Psychosomatik
PPV	Positive Predictive Value (Precision)
PZN	Pharmazentralnummer
NPV	Negative Predictive Value
RF	Random Forest
SGB	Sozialgesetzbuch
XAI	Explainable AI (engl.) / Erklärbare KI (dt.)
OTC	Over The Counter

Begriffs-/Synonymverzeichnis

Begriff	Synonym / Übersetzung / Erklärung
Activation Function	Aktivierungsfunktion (für die Weitergabe von Signalen in Neuronalen Netzen)
Bootstrapping	Methode, bei der wiederholt Stichproben mit Zurücklegen aus einem Datensatz gezogen werden (um die Genauigkeit von Schätzungen oder Modellen zu beurteilen)
Confusion Matrix	Konfusionsmatrix oder Fehlermatrix (dt.): Vier-Felder-Tafel mit Gegenüberstellung der vorhergesagten und tatsächlichen Werte eines Klassifizierungsalgorithmus
Dimensionality Reduction	Dimensionsreduktion; Reduktion der Anzahl an separat berücksichtigten Variablen in einer Analyse
Dropout	Hier: „Ausschaltung“ von Neuronen im Neuronalen Netzwerk
Feature Space	Merkmalsraum, in dem die Merkmale (Features) die Dimensionen des Raums bestimmen
Loss Function	Verlustfunktion, Kostenfunktion; ordnet als mathematische Funktion den Abweichungen vorhergesagter von den tatsächlichen Werten „Verluste“ zu, die dann im Zuge der Optimierung eines Vorhersagemodells minimiert werden sollen
Odds	Chance; Quotient aus Ereignissen und komplementären Ereignissen (z. B. Verstorbene / nicht Verstorbene)
Odds Ratio	Chancenverhältnis; Quotient der Odds aus zwei Gruppen (z. B. Verstorbene / nicht Verstorbene bei Rauchern geteilt durch die entsprechende Odds bei Nichtrauchern)
Overfitting	Überanpassung eines Modells an die verwendeten Trainingsdaten, die nicht generalisiert werden können
Principal Component Analysis	Hauptkomponentenanalyse; Verfahren der Dimensionsreduktion
Reinforcement Learning	Bestärkendes Lernen (s. a. Kapitel 2.1)
Sensitivity	Sensitivität / Recall; Anteil der vorhergesagten/erkannten positiven Fälle an allen tatsächlich positiven Fällen (s. a. Kapitel 3.1)
Specificity	Spezifität; Anteil der vorhergesagten/erkannten negativen Fälle an allen tatsächlich negativen Fällen (s. a Kapitel 3.1)
Supervised Learning	Überwachtes Lernen (s. a Kapitel 2.1)
Trainingsbatch	Teilmenge eines Datensatzes für das Trainieren von Modellen
Unsupervised Learning	Unüberwachtes Lernen (s. a. Kapitel 2.1)
Up- und Downsampling	Verfahren zur Erhöhung bzw. zur Reduktion der Anzahl von Beobachtung mit bestimmten Merkmalsausprägungen in den Daten (s. a Kapitel 2.6)
Zielvariable	Abhängige oder zu erklärende Variable, Prognosevariable, Regressand (bei Regressionsanalysen); Target variable oder Outcome (engl.)

Einleitung

In allen Lebensbereichen hat die Menge an digital verfügbaren Informationen in den letzten Jahren erheblich zugenommen. Dies gilt auch für den Gesundheitsbereich. Gesundheitsdaten werden dabei an unterschiedlichen Stellen zu unterschiedlichen Zwecken von unterschiedlichen Beteiligten erfasst und sind dann über mehr oder minder lange Zeiträume für bestimmte Zwecke von zumeist eingeschränkten Personenkreisen nutzbar. Ein – zumindest grundsätzlich – unstrittiges und erstrebenswertes Ziel der Nutzung entsprechender Daten besteht darin, mit Analysen der Daten zur Verbesserung der Gesundheit ganz allgemein oder bezogen auf bestimmte Personengruppen und Gesundheitsbereiche beizutragen. Neben eher klassischen Auswertungsmethoden werden hierfür vermehrt auch Techniken und Methoden der Künstlichen Intelligenz (KI) genutzt.

Routinedaten der gesetzlichen Krankenversicherung

Ein nicht unwesentlicher Teil gesundheitsbezogener Daten wird in Deutschland primär zu Abrechnungszwecken und zur Prüfung und Abwicklung von Versicherungsleistungen der gesetzlichen Krankenversicherung (GKV) erfasst, über die in Deutschland die gesundheitliche Versorgung von mehr als 85 % der Bevölkerung abgesichert ist. Um die Abrechnung einer Vielzahl an Leistungserbringern (wie Arztpraxen, Krankenhäusern und Apotheken) mit einer weiterhin hohen zweistelligen Zahl an gesetzlichen Krankenkassen zu erleichtern und bestimmte Berichtspflichten der GKV-Kassen zu ermöglichen, sind Formate und Umfänge der Datenübermittlung in vielen Bereichen bundesweit einheitlich geregelt. So verfügen alle GKV-Kassen über vergleichbare Informationen zu den jeweils bei ihnen versicherten Personen, die nachfolgend im Weißbuch kurz als Routinedaten bezeichnet werden. Routinedaten einzelner Krankenkassen wurden bereits in vergangenen Jahren auch für gesundheitswissenschaftliche Auswertungen und Analysen genutzt. Noch innerhalb des Jahres 2025 sollen wesentliche Daten aller GKV-Kassen zu mehreren Erhebungsjahren im Forschungsdatenzentrum Gesundheit (FDZ) für bestimmte Analysen verfügbar sein, womit dann sehr umfangreiche und strukturierte Daten zu einem Großteil der Bevölkerung Deutschlands an einer Stelle für Analysen genutzt werden können (Stand Juli 2025).

Ziele und Inhalte des Weißbuchs

Das vorliegende Weißbuch konnte im Rahmen eines durch den Innovationsausschuss beim Gemeinsamen Bundesausschuss unter dem Förderkennzeichen 01VSF20014 geförderten Forschungsprojektes KI-THRUST erstellt werden, wofür an dieser Stelle ganz herzlich gedankt sei. Das Weißbuch verfolgt als maßgebliches Resultat des Forschungsprojektes zunächst zwei grundlegende Ziele.

1. Zum einen sollen praxisrelevante Einblicke in die Strukturen und Merkmalsumfänge von GKV-Routinedaten gegeben werden.
2. Zum anderen sollen Methoden der Künstlichen Intelligenz vorgestellt und erläutert werden, die mit entsprechenden Daten für die Vorhersage von Ereignissen genutzt werden können.

Während das erste Ziel auf Leserinnen und Leser ausgerichtet ist, die zuvor nicht oder wenig mit GKV-Routinedaten gearbeitet haben, ist das zweite Ziel vorrangig auf diejenigen ausgerichtet, die einen Einstieg in die Nutzung von Methoden der Künstlichen Intelligenz mit ebensolchen Daten zur Vorhersage von Ereignissen im Sinne einer Risikoprädiktion suchen. Diese beiden grundlegenden Ziele werden auf allgemeiner Ebene vorrangig in Kapitel 1 und 2 des Weißbuchs adressiert.

Sowohl traditionelle Verfahren zur Vorhersage von Ereignissen wie logistische Regressionsmodelle als auch Vorhersagen basierend auf KI-Modellen müssen hinsichtlich ihrer Vorhersagegüte beurteilt werden, um in der Praxis zwischen mehr oder weniger geeigneten Prädiktionsmodellen unterscheiden zu können. Hierfür existiert eine Reihe von Kennwerten mit teils vielfältigen Bennungen, die in Kapitel 3 erklärt werden. Hingewiesen sei an dieser Stelle darauf, dass nach den Erfahrungen bei der Erstellung des Weißbuchs eine Reihe von Missverständnissen zwischen eher traditionell und KI-orientierten Datenanalysten schlicht aus unterschiedlichen Benennungen derselben Sachverhalte resultieren können.

In den nachfolgenden Kapiteln 4 bis 8 des Weißbuchs werden die Aufbereitung von Routinedaten, die Umsetzung von konventionellen und KI-basierten Modellrechnungen sowie der Vergleich der Prädiktionsgüte der Modelle schrittweise erläutert. Es folgen ergänzende Analysen zur Erklärbarkeit, zu Fehlklassifikationen und zur Übertragbarkeit von Vorhersagen auf andere Beobachtungszeiträume. Die Darstellungen in diesen Kapiteln basieren maßgeblich auf Auswertungen, für die auf Routinedaten zu insgesamt rund 1,4 Millionen BKK-Versicherten mit mindestens einem Krankenhausaufenthalt in den Jahren von 2015 bis 2020 zurückgegriffen werden konnte. Mit den Darlegungen sollen praktische Vorgehensweisen, Ergebnisse und deren Bewertungen präsentiert werden, die als exemplarische Beispiele von Aufbereitungen und Analysen „realer“ Routinedaten Anhaltspunkte und Anregungen für Vorgehensweisen bei eigenen Analysen liefern. Das Kapitel 9 widmet sich schließlich den Voraussetzungen und Möglichkeiten einer Anwendung von KI-Verfahren in der Routineversorgung und bei Krankenkassen.

Im Sinne der eingangs genannten grundlegenden Ziele hoffen wir als Autorinnen und Autoren, mit dem Weißbuch sowohl zum Verständnis von GKV-Routinedaten und den damit bestehenden Analysemöglichkeiten als auch zum Verständnis von KI-Methoden und deren Anwendung mit Rückgriff auf Routinedaten beitragen zu können.

1 Routinedaten bei gesetzlichen Krankenkassen

Routinedaten bei gesetzlichen Krankenkassen beinhalten vorrangig Informationen, die zur Prüfung und Vergütung von Gesundheitsleistungen erforderlich sind.

Bei größeren Krankenkassen sind dies Daten,

- *die zu mehreren hunderttausend Personen vorliegen,*
- *die sich längsschnittlich verknüpfbar auf Zeiträume über mehrere Jahre beziehen und die, neben grundlegenden Informationen, auch*
- *umfangreiche Informationen zur Inanspruchnahme einer Vielzahl medizinischer Leistungsbereiche umfassen (z. B. Arzneiverordnungen, ambulante und stationäre Behandlungen).*

Der Datenaustausch zwischen Krankenkassen und Leistungserbringern ist grundlegend im SGB V gesetzlich geregelt. Dadurch ist der Umfang der Daten, die bei den Krankenkassen vorliegen, weitgehend standardisiert.

Eines der Anliegen des Weißbuchs ist es, eine möglichst konkrete Vorstellung zu den bei gesetzlichen Krankenkassen regelmäßig verfügbaren Daten zu vermitteln, die aufgrund ihrer fortlaufenden Erfassung auch als Routinedaten bezeichnet werden können. In der Regel werden diese Daten bei Krankenkassen zu Abrechnungszwecken erhoben, also um Vergütungen der Krankenkassen an Leistungserbringer (z. B. Arztpraxen, Krankenhäusern oder Apotheken) zu begründen. Dabei beinhalten die Daten eine Vielzahl an gesundheitsrelevanten Informationen. Primär lassen sich zunächst die abgerechneten Inanspruchnahmen der gesundheitlichen Versorgung darstellen. Zugleich sind mit den Daten jedoch auch Aussagen zu Häufigkeiten und zu Behandlungen von Erkrankungen möglich (Schubert, Köster, Küpper-Nybelen, & Ihle, 2008).

Potenzial von Routinedaten

Das Potenzial der Routinedaten ergibt sich aus der Art und Weise wie diese Daten zustande kommen (Neubauer, Zeidler, Lange, & Graf von der Schulenburg, 2017; Schubert et al., 2008; Swart & Ihle, 2008): Vor allem bei größeren Krankenkassen sind Daten zu ausgesprochen großen Populationen verfügbar, die oftmals mehrere hunderttausend Personen umfassen. Mögliche Untersuchungspopulationen bewegen sich damit in einem Größenbereich, der im Rahmen wissenschaftlicher Studien bei Primärerhebungen nicht erreichbar ist. Routinedaten können so einen sonst nicht verfügbaren Einblick

in die Versorgungsrealität zu sehr großen Populationen mit entsprechenden Differenzierungsmöglichkeiten liefern.

Darüber hinaus gibt es bei Routinedaten im Vergleich zu Primärdaten weniger Verzerrung aufgrund von (Nicht-)Teilnahme bestimmter Populationen. So können mithilfe von Routinedaten auch Personengruppen analysiert werden, die normalerweise schwierig oder gar nicht für Primärforschung zur Verfügung stehen, wie z. B schwerstkranke oder demente Personen (Neubauer et al., 2017). Die Vollständigkeit von Routinedaten erstreckt sich häufig auch über mehrere Jahre, was Langzeitanalysen ermöglicht.

Routinedaten bei anderen Versicherungsträgern

Im Rahmen dieses Weißbuchs soll insbesondere auf solche Routinedaten eingegangen werden, die bei gesetzlichen Krankenversicherungen vorliegen. Darüber hinaus existieren gesundheitsbezogene Routinedaten aber auch an anderen Stellen, z. B bei der privaten Krankenversicherung, bei der gesetzlichen Pflegeversicherung (SGB XI; s. Schwinger et al. (2018) für eine beispielhafte Nutzung), den Renten- und Unfallversicherungen oder im Rahmen der amtlichen Statistik. Eine Betrachtung all dieser Datenquellen würde jedoch den Rahmen des Weißbuchs sprengen, welches einen fokussierten Überblick zu Routinedaten bei gesetzlichen Krankenkassen geben möchte.

Gesetzliche Regelungen zum Datenaustausch

In Deutschland sind mehr als 70 Mio. Menschen bei einer der knapp 100 Krankenkassen der gesetzlichen Krankenversicherung (GKV) krankenversichert. An der gesundheitlichen Versorgung der GKV-Versicherten sind dabei unter anderem knapp 2.000 Krankenhäuser, knapp 20.000 Apotheken und mehr als 150.000 ambulant tätige Ärzte/Ärztinnen sowie mehr als 30.000 Psychologische Psychotherapeuten/Psychotherapeutinnen (für Erwachsene und Kinder- und Jugendliche) beteiligt (s. Statistisches Bundesamt; Statistik des Deutschen Apothekerverbands; Ärztestatistik 2021 der Bundesärztekammer). Um Abrechnungen dieser großen Zahl an Leistungserbringern mit den jeweils zuständigen Krankenkassen zu ermöglichen, sind eine Vielzahl an Festlegungen notwendig. Der Datenaustausch zwischen Leistungserbringern (also den Apothekern/Apothekerinnen, Therapeuten/Therapeutinnen, Krankenhäusern usw.) und den gesetzlichen Krankenkassen ist dabei grundlegend im Fünften Buch „Gesetzliche Krankenversicherung“ des Sozialgesetzbuchs (SGB V) geregelt und bereits dadurch weitgehend standardisiert. Details der Datenübermittlung werden zwischen Krankenkassen und Leistungserbringern vereinbart und zumeist in sogenannten Technischen Anlagen (TA) beschrieben (vgl. gkv-datenaustausch.de). Durch diese Regelungen erhalten alle gesetzlichen Krankenkassen eine Vielzahl an Daten aus unterschiedlichen Bereichen des Gesundheitssystems in einem einheitlichen Format. Im Hinblick auf die wesentlichen Leistungsbereiche sollten so alle gesetzlichen Krankenkassen in Deutschland – bezogen auf die jeweils bei ihnen versicherten Personen – über gleichartig differenzierte Daten zu Inanspruchnahmen der gesundheitlichen Versorgung verfügen.

GKV-Leistungsbereiche

Zu den relevantesten Leistungsbereichen der gesundheitlichen Versorgung, auf deren Daten nachfolgend im Detail eingegangen werden soll, zählen die in Tabelle 1-1 gelisteten Bereiche. Darüber hinaus liegen bei den Krankenkassen mit Stammdaten und Versicherungshistorien weitere grundlegende Informationen über die Versicherten vor.

Tabelle 1-1. Wesentliche Leistungsbereiche der gesundheitlichen Versorgung

Leistungsbereich	Informationen
Ambulante ärztliche Versorgung	Leistungen, die von niedergelassenen Vertragsärzten/Vertragsärztinnen im Rahmen der ambulanten Versorgung üblicherweise in Praxen erbracht werden
Krankenhausbehandlungen	Leistungen, die im Krankenhaus erbracht werden
Arzneimittel	Arzneimitteln, die von Ärzten/Ärztinnen verordnet und typischerweise von Apotheken ausgegeben werden
Heilmittel	Daten zu Physikalischer Therapie, Ergotherapie, Ernährungstherapie, Logopädie, Podologie, die die Krankenbehandlung sichern bzw. einer Behinderung vorbeugen
Hilfsmittel	Produkte, die die Krankenbehandlung sichern bzw. einer Behinderung vorbeugen oder sie ausgleichen
Arbeitsunfähigkeit und Krankengeld	Erkrankungsbedingte Arbeitsunfähigkeiten und ggf. im Sinne von Entgeltfortzahlungen von der Krankenkasse gezahltes Krankengeld
Rehabilitation	Ambulante und stationäre Rehabilitationsmaßnahmen
Zahnärztliche Behandlungen	Leistungen, die von niedergelassenen Vertragszahnärzten/Vertragszahnärztinnen erbracht werden

Datenorganisation bei den Krankenkassen

In welcher Form die Daten bei den Krankenkassen gespeichert sind, also die genaue Datenorganisation und -struktur, kann von Krankenkasse zu Krankenkasse unterschiedlich sein. Die Strukturen der Datenhaltung sind i.d.R. jedoch bei unterschiedlichen Krankenkassen ähnlich. Die Daten liegen bei den Krankenkassen typischerweise in relationalen Datenbanken mit zumeist mehreren Tabellen für Daten aus jeweils einem Leistungsbereich vor, wie dies exemplarisch in Abbildung 1-1 dargestellt ist. Verknüpfungen zwischen den Tabellen sind über Identifikationsmerkmale möglich, die typischerweise einzelne Versicherte und einzelne Abrechnungsfälle kennzeichnen.

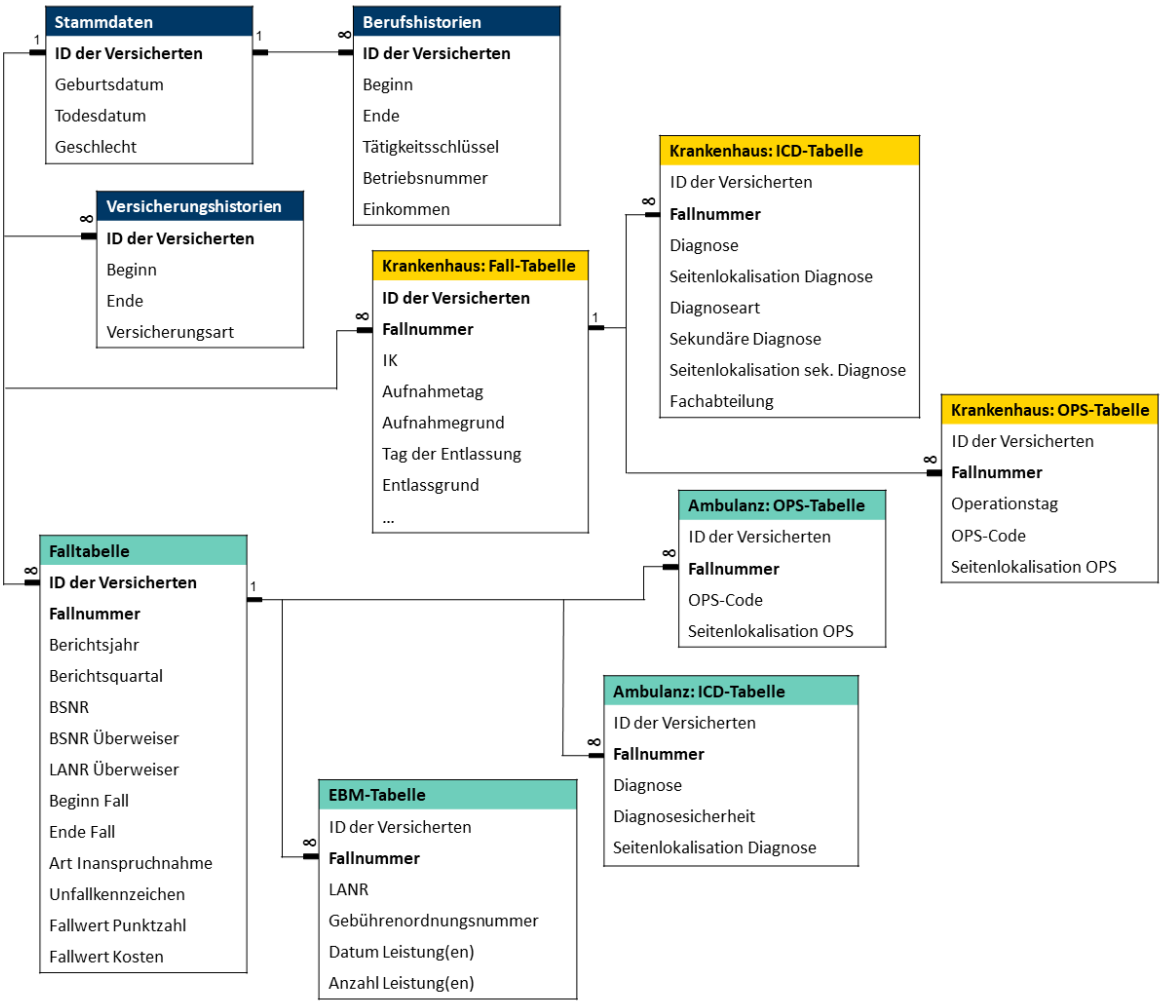

Abbildung 1-1. Beispielhafte Darstellung der Datenorganisation in einer relationalen Datenbank bei einer Krankenkasse

Datenschutz und Pseudonymisierung

Die Routinedaten, die bei den Krankenkassen vorliegen, beinhalten immer personenbezogene Daten, also identifizierende Informationen über die Versicherten und die Leistungserbringer, um die Abrechnung zu ermöglichen. Zur eindeutigen Identifikation von Versicherten und Leistungserbringern dienen persönlich zugeordnete Kennnummern, wie die Krankenversichertennummer für die Versicherten und die lebenslange Arztnummer (LANR) oder die Betriebsstättennummer (BSNR) für die Leistungserbringer (s. Abschnitt 1.8 „Kennzeichnungen von Personen und Einrichtungen im Gesundheitssystem“).

Aufgrund des Personenbezugs und des teilweise hochsensiblen Inhalts der Krankenkassenroutinedaten müssen bei der Nutzung der Daten Vorsichtmaßnahmen getroffen werden, die einen Missbrauch verhindern. So werden die Daten i.d.R. pseudonymisiert verarbeitet, d. h. personenidentifizierende Merkmale wie die Krankenversichertennummer sind durch eine andere Kennung ersetzt, so dass einzelne Personen nicht mehr direkt identifizierbar sind, aber dennoch eine Verknüpfung der Daten aus unterschiedlichen Leistungsbereichen (also zwischen den verschiedenen Tabellen) möglich ist.

Auch pseudonymisierte Daten können für Forschungszwecke nicht einfach unmittelbar genutzt werden. Routinedaten von Krankenkassen zählen im juristischen Sinne zu den sogenannten Sozialdaten, die dem Sozialdatenschutz unterliegen und zunächst nur für ihren vorgesehenen Verwendungszweck verarbeitet werden dürfen (vgl. § 67 SGB X). Im Rahmen des SGB X ist darüber hinaus geregelt, wie

und unter welchen Voraussetzungen dennoch eine Verwendung für die Forschung möglich ist. Zum einen können die Daten genutzt werden, wenn der Betroffene explizit und aus freien Stücken in die Nutzung einwilligt. Dies ist aber nur für Forschungsvorhaben zu einem eingegrenzten Personenkreis praktikabel, da alle Versicherten persönlich kontaktiert werden müssten. In § 75 SGB X wird zusätzlich geregelt, unter welchen Bedingungen eine Übermittlung der Daten zu Forschungszwecken möglich ist, ohne die Einwilligung der betroffenen Personen einzuholen. Dies ist allerdings nur zulässig, wenn die Daten für das Forschungsvorhaben unverzichtbar sind und das öffentliche Interesse an der Forschung so hoch ist, dass es das private Interesse der Betroffenen übersteigt. Für die Übermittlung muss eine Genehmigung durch die für die Krankenkassen zuständigen Aufsichtsbehörden auf Landes- und/oder Bundesebene erstellt werden.

Bei der Nutzung der Daten zu Forschungszwecken müssen ebenfalls Vorkehrungen getroffen werden, um die Identifikation der Personen zu verhindern. Eine Möglichkeit stellt das komplette Entfernen des Personenbezugs – sprich die Löschung sämtlicher Klardaten und Identifikatoren – dar, um die Daten zu anonymisieren. Je nach Forschungsvorhaben und Art der Daten ist dies allerdings nicht möglich bzw. nicht sinnvoll, weil eine Zuordnung der Daten aus unterschiedlichen Leistungsbereichen und unterschiedlichen Jahren zu einer Person dann nicht mehr möglich ist. Die Daten werden daher i.d.R. vor der Weitergabe an die Forschenden pseudonymisiert, d. h. der Name und andere Identifikationsmerkmale (wie Adresse, Identifikationsnummern) werden durch ein Kennzeichen ersetzt, um die Identifikation der Betroffenen zu verhindern bzw. zu erschweren (dies betrifft sowohl die Versicherten als auch die Leistungserbringer und Institutionen). Manche Daten müssen auch vergröbert werden, um eine Identifikation zu verhindern. So wird z. B statt dem genauen Geburtsdatum nur das Geburtsjahr weitergegeben. Auch die Postleitzahl wird häufig in reduziertem Umfang zur Verfügung gestellt (z. B auf die ersten drei Ziffern gekürzt), um eine Identifikation von Personen zu vermeiden, aber eine Auswertung nach Regionen dennoch zu ermöglichen.

Gliederung und Inhalte nachfolgender Abschnitte

In den nachfolgenden Abschnitten soll auf einzelne Leistungsbereiche, aus denen Daten bei den Krankenkassen verfügbar sind, näher eingegangen werden (Kapitel 1.1 bis 1.6). Darin gibt es nach einer allgemeinen Einführung in die Besonderheiten des jeweiligen Datenbereichs Informationen zu Struktur und Merkmalsumfang der Daten, sowie Informationen zum Datenvolumen, den Übermittlungsweg von Leistungserbringer zu Krankenkasse und Nutzungshinweise. Es werden vor allem die Daten hervorgehoben, die sich typischerweise zu Forschungszwecken eigenen. Eine umfassende Auflistung über sämtliche Daten, die bei den Krankenkassen vorliegen sollten, wird nicht erstellt, da dies den Rahmen des Weißbuchs sprengen würde. Für Interessierte lässt sich eine solche Auflistung aber anhand der sogenannten Technischen Anlagen (TA) der Richtlinien des Datenaustauschverfahrens erstellen.

Im Anschluss werden die wichtigsten Klassifikationssysteme (z. B für Krankheiten oder Arzneimittel) vorgestellt, die im Gesundheitswesen relevant sind (Kapitel 1.7), sowie Kennzeichnungen für Personen und Einrichtungen (Kapitel 1.8). Im letzten Kapitel sind Besonderheiten von Krankenkassenroutinedaten zusammengefasst, die die Nutzung der Daten zu Forschungszwecken beeinflussen können (Kapitel 1.9).

1.1 Stammdaten, Versicherungs- und Berufshistorien

In den Stammdaten sind grundlegende Informationen zu den Versicherten gespeichert, wie Name, Geburtsjahr und Adresse, sowie Versicherungszeiten.
Man kann zwei Gruppen von Versicherten unterscheiden, Mitglieder und Familienversicherte, über die unterschiedlich viele Informationen gespeichert sind.

Um ihre Aufgaben zu erfüllen, müssen die Krankenkassen notwendigerweise einige grundlegende Angaben zu den Versicherten speichern, die häufig als Stammdaten bezeichnet werden. Dazu zählen beispielsweise Name, Adresse und Geburtsdatum der Versicherten, sowie Informationen zum Versicherungsstatus.

Die Versicherten einer gesetzlichen Krankenkasse können in verschiedene Gruppen aufgeteilt werden. Zunächst kann man die Mitglieder einer gesetzlichen Krankenkasse von den Familienversicherten unterscheiden. Zu den Mitgliedern zählen alle Personen für die Beitragszahlungen zur Krankenversicherung anfallen. Dazu zählen insbesondere pflichtversicherte Arbeitnehmer, die die größte Gruppe darstellen (Bundesministerium für Gesundheit, 2021). In Deutschland besteht eine Versicherungspflicht in der gesetzlichen Krankenversicherung für alle Arbeitnehmer, deren Jahresgehalt unter der Versicherungspflichtgrenze liegt (Stand 2025: 73.800 €). Weitere Mitgliedsgruppen sind freiwillig Versicherte mit einem Einkommen oberhalb der Versicherungspflichtgrenze, Arbeitslosengeldempfänger und Rentner (§5ff SGB V). Ehepartner und Kinder von Mitgliedern, die über kein eigenes Einkommen verfügen, sind in der GKV i.d.R. als Familienversicherte beitragsfrei über die Mitglieder mitversichert (z. B Ehe-/Lebenspartner, Eltern; §10 SGB V). Die Familienversicherten sind daher kassenintern stets mit den Mitgliedern verbunden, über die sie versichert sind. Kinder können i.d.R. bis zum 18 Lebensjahr mitversichert werden bzw. bis zum 25. Lebensjahr, sofern sich das Kind bis zu diesem Alter in Ausbildung oder Studium befindet.

1.1.1 Struktur und Merkmalsumfang der Daten

In den Stammdaten sind grundlegende Informationen zu den versicherten Personen gespeichert (s. Tabelle 1-2) wie Vor- und Nachname, Geburtsdatum, Geschlecht, Wohnort mit Postleitzahl, ggf. Staatsangehörigkeit und Familienstand. Zu den verschiedenen Versichertengruppen sind unterschiedlich viele Informationen in den Stammdaten vorhanden (wohingegen die leistungsbezogenen Daten für alle Gruppen identisch erfasst werden). Zu Mitgliedern sind beispielsweise auch Informationen zu beitragspflichtigen Einkünften bzw. zur Beitragshöhe und -einstufung vorhanden, die es bei den Familienversicherten nicht gibt, da diese beitragsfrei versichert sind und i.d.R. kein Einkommen haben.

Versicherungshistorien

Für alle Versicherten, also Mitglieder und Familienversicherte, müssen die Krankenkassen ein Versicherungsverzeichnis führen (§288 und §289 SGB V). Neben den bisher erläuterten personenbezogenen Daten dokumentieren die Krankenkassen daher zu allen Versicherten die Versicherungszeiten, i.d.R. mit tagesgenauem Beginn- und Enddatum der Versicherung. Typischerweise wird auch ein Austrittsgrund bei Beendigung der Versicherung erfasst. Diese Informationen erlauben, die Versicherungshistorien der Versicherten zu verfolgen und sind somit für die meisten Auswertungen wichtig, insbesondere für Morbiditätsschätzungen. Mit ihrer Hilfe kann die Bezugsgruppe (der Nenner) entsprechend angepasst werden. So kann z. B. nachvollzogen werden, ob für Personen keine ambulanten Leistungen vorliegen, weil die Personen nicht bei einem Vertragsarzt in Behandlung waren oder weil sie nicht (mehr) bei der Krankenkasse versichert waren.

Informationen zur beruflichen Tätigkeit

Zu den Arbeitnehmenden liegen zusätzlich Informationen zur Tätigkeit vor, da die Arbeitgeber den Krankenkassen für alle Arbeitnehmenden Angaben zur Tätigkeit melden müssen (SGB IV §28a). Grundlage hierfür ist der Tätigkeitsschlüssel[1] der Bundesagentur für Arbeit. Der Tätigkeitsschlüssel beinhaltet neben Informationen zum Beruf auch Informationen zur höchsten erreichten Schulbildung und beruflichen Bildung, sowie der Vertragsform der Beschäftigten. Die Aktualität und Validität dieser Angaben hängt maßgeblich von der Erfassung und den Meldungen durch die Arbeitgeber ab. Auch Angaben zum Arbeitgeber selbst werden gespeichert (in Form einer Betriebsnummer, über die der Arbeitgeber stets auch einer Branche zugeordnet sein sollte). Außerdem muss der Krankenkasse das monatliche Erwerbseinkommen der Arbeitsnehmenden bekannt sein (bis zur Beitragsbemessungsgrenze, Stand 2024: 62.100 €), um den Krankenkassenbeitrag zu berechnen.

Tabelle 1-2. Stammdaten, Versicherungs-, Berufs- und Wohnhistorien – Struktur und Merkmale

Merkmal	Erläuterung
Stammdaten (ein Eintrag pro Person)	
Versicherten-ID*	Identifikationsnummer der Versicherten (pseudonymisiert)
Geburtsdatum	Für Forschungszwecke i.d.R. reduziert auf das Geburtsjahr
Todesdatum	Todesdatum der Versicherten, sofern zutreffend
Geschlecht	Geschlecht der Versicherten
Versicherungshistorien (einer bis viele Einträge pro Person)	
Versicherten-ID*	Identifikationsnummer des Versicherten (pseudonymisiert)
Beginn	Beginn der Versicherungszeit
Ende	Ende der Versicherungszeit (ggf. Datum, das weit in der Zukunft liegt bei noch andauernder Versicherung)
Versicherungsart	Mitglied (berufstätig, berentet, arbeitslos)/Familienversichert
Wohnhistorien (einer bis viele Einträge pro Person)	
Versicherten-ID*	Identifikationsnummer des Versicherten (pseudonymisiert)
Beginn	Datum ab dem die Anschrift/Adresse gilt
Ende	Datum bis zu dem die Anschrift/Adresse gilt
Straße	Anschrift der Versicherten (zu Forschungszwecken i.d.R. reduziert auf Regionen oder dreistellige Postleitzahlbereiche)
Postleitzahl	
Wohnort	
Berufshistorien (einer bis viele Einträge pro Arbeitnehmer*in)	
Versicherten-ID*	Identifikationsnummer des Versicherten (pseudonymisiert)
Beginn	Beginn der Beschäftigung
Ende	Ende der Beschäftigung
Tätigkeitsschlüssel	Kennziffer für die Tätigkeit auf Basis des Tätigkeitsschlüssel der Bundesagentur für Arbeit
Betriebsnummer	Betriebsnummer des Arbeitsgebers
Einkommen	Monatliches Erwerbseinkommen bis zur Beitragsbemessungsgrenze

1 abrufbar unter: https://www.arbeitsagentur.de/betriebsnummern-service/taetigkeitsschluessel, Stand 01/2025

1.1.2 Datenvolumen, Übermittlung, Verfügbarkeit sowie Nutzungshinweise

Die Eintragungen in den Stammdaten sind übersichtlich, da hier genau ein Eintrag pro Person existiert mit feststehenden (also konstanten bzw. aktuellen) Merkmalen. Die Eintragungen in den Tabellen zu Versicherungs-, Wohn- und Berufshistorien hingegen können deutlich komplexer werden, besonders wenn eine längere Zeitspanne betrachtet wird, da dann bei jedem Statuswechsel ein neuer Eintrag erfolgt. Informationen über Übermittlung, Umfang und Verfügbarkeit der Daten in den verschiedenen Datentabellen sind in Tabelle 1-3 zusammengefasst.

Tabelle 1-3. Stammdaten, Versicherungs-, Berufs- und Wohnhistorien – Übermittlung und Verfügbarkeit der Daten

Datentabelle	Erläuterungen
Stammdaten	Von Krankenkassen selbst erfasst, obligat pro Person genau ein Tabelleneintrag vorhanden, i.d.R. zeitnah verfügbar
Versicherungshistorien	Von Krankenkassen selbst erfasst, obligat pro Person mindestens ein Tabelleneintrag/Intervall vorhanden, selten auch sehr viele Intervalle mit zwischenzeitlichen Unterbrechungen der Versicherung möglich, i.d.R. zeitnah verfügbar
Wohnhistorien	Von Krankenkassen selbst erfasst, obligat pro Person mindestens ein Tabelleneintrag/Intervall vorhanden, bei häufigen Wohnortwechseln auch viele Intervalle möglich, nicht von allen Kassen gepflegt, meldeabhängig verfügbar
Berufshistorien	Von den Arbeitgebern an die Krankenkassen gemeldet, obligat zu jedem sozialversicherungspflichtigen Beschäftigungsintervall mit Tätigkeitsschlüssel sowie Angabe zur Branche des Arbeitgebers verfügbar, teils auch im Sinne von Mitgliedschaftshistorien zu nicht berufstätigen Mitgliedern, bei gleichzeitiger Beschäftigung durch mehrere Arbeitgeber auch überlappend dokumentierte Zeiträume möglich, selten auch sehr viele Intervalle möglich

Aus den in den Stammdaten verfügbaren Variablen ergeben sich verschiedene Nutzungsmöglichkeiten (Grobe & Ihle, 2014): So können die Daten genutzt werden, um geschlechts-, altersgruppen- oder regionalabhängige Analysen zu berechnen. Aus den Daten, die über Arbeitnehmer/-innen vorliegen (monatliches Erwerbseinkommen bis zur Beitragsbemessungsgrenze, ausgeübter Beruf, höchster Schulabschluss und höchste berufliche Bildung) lässt sich außerdem der soziale Status approximieren, der für die empirische Sozialforschung und Epidemiologie häufig relevant ist. Hier muss allerdings berücksichtigt werden, dass nur das beitragspflichtige Einkommen erfasst wird, andere Vermögenswerte sowie das Einkommen eines möglichen Partners aber außen vorbleiben. Durch den Bezug von Familienversicherten zu den Hauptmitgliedern lassen sich z. B. ggf. Verknüpfungen zwischen Müttern und Neugeborenen herstellen, sofern Neugeborene über die Mutter versichert sind.

1.2 Ambulante ärztliche Versorgung

Daten aus der ambulanten ärztlichen Versorgung werden i.d.R. über die kassenärztlichen Vereinigungen an die Krankenkasse gemeldet. Sie enthalten Informationen über die behandelnden Ärzte/Ärztinnen, die erbrachten Leistungen und zu Diagnosen. Von einer Praxis werden die Daten einmal pro Quartal als ein „Fall" gemeldet bzw. abgerechnet, auch wenn Versicherte mehrere Kontakte innerhalb des Quartals hatten.

Ambulante ärztliche (und psychotherapeutische) Behandlungen bilden einen wesentlichen Leistungsbereich der Krankenversicherung. Die ambulante ärztliche Versorgung der gesetzlich Versicherten in Deutschland und die damit verbundene Abrechnung wird von aktuell 17 kassenärztlichen Vereinigungen (KV) verantwortet (§§72ff. SGB V). Die KVen haben regionale Zuständigkeiten (entsprechend der 16 Bundesländer mit Ausnahme von Nordrhein-Westfalen, das in die KV Nordrhein und KV Westfalen-Lippe geteilt ist). Ärzte/Ärztinnen können nur dann an der regulären ambulanten Versorgung von gesetzlich Versicherten teilnehmen und zulasten der gesetzlichen Krankenversicherung abrechnen, wenn sie eine Zulassung bzw. Ermächtigung der KVen haben. Auch die Abrechnung der Leistungen läuft über die KVen, d. h. die Ärzte/Ärztinnen rechnen ihre Leistungen mit der regionalen KV ab, die wiederum das Geld von den Krankenkassen erhält. In §295 SGB V ist geregelt, welche Daten die KVen zur Abrechnung auf elektronischem Wege an die gesetzliche Krankenversicherung weiterzugeben haben.

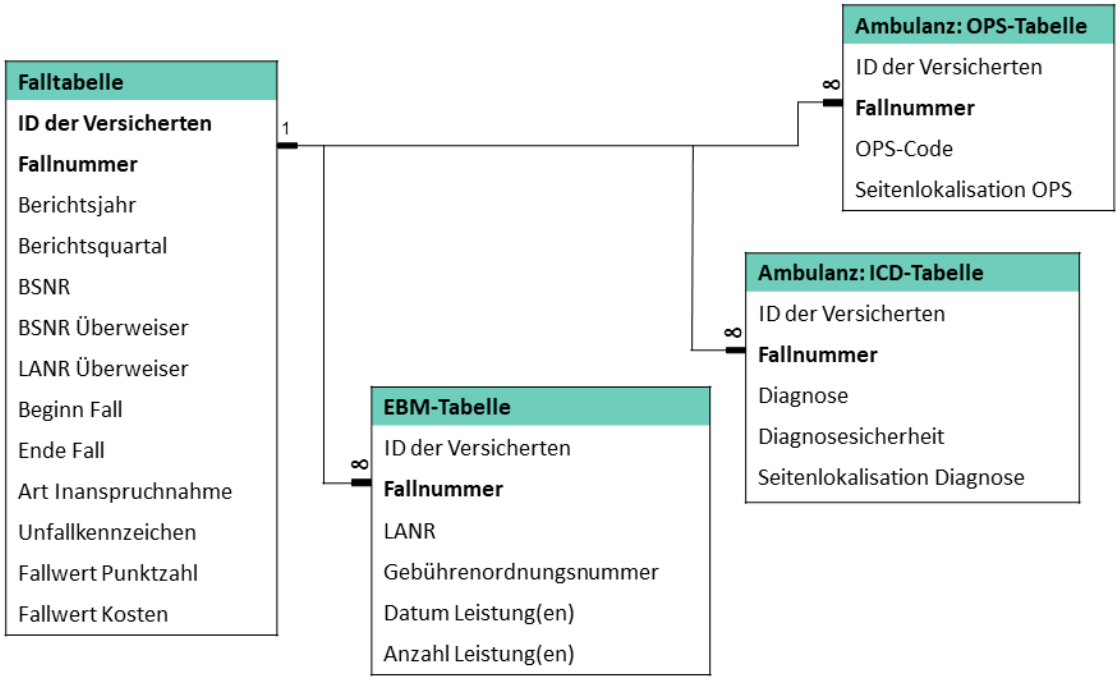

Abbildung 1-2. Schematische Darstellung der ambulanten Datentabellen

1.2.1 Struktur und Merkmalsumfang der Daten

Einen Überblick über die Datenstruktur der ambulanten Datentabellen bietet Abbildung 1-2. In den bei Krankenkassen gespeicherten Daten aus dem ambulanten ärztlichen Bereich finden sich Informationen über die behandelnden Ärzte/Ärztinnen, Art und Umfang erbrachter Leistungen und den ermittelten Diagnosen (Neubauer et al., 2017; Swart, Ihle, Gothe, & Matusiewicz, 2014).

Abrechnungsfall

Die zentrale Beobachtungseinheit im ambulanten Bereich ist der Abrechnungsfall, wobei ein Abrechnungsfall alle Behandlungen einer Person bei einem Arzt oder einer Ärztin innerhalb eines Quartals umfasst. Jeder Abrechnungsfall hat eine eindeutige Nummer und zu jedem Fall wird die Versichertenkennung, die Kennung des behandelnden Arztes/der behandelnden Ärztin (s. Kapitel 1.8) und ggf. die Kennung des überweisenden Arztes/der überweisenden Ärztin gespeichert. Für jeden Abrechnungsfall kann eine (nahezu) beliebige Anzahl an Diagnosen angegeben werden (s. Kapitel 1.7.1), wobei i.d.R. zumindest eine Diagnose angegeben werden muss. Die Diagnosen sind ohne Datum erfasst. Zu ihnen wird jeweils auch die Sicherheit der Diagnose angegeben (G = gesicherte Diagnose, A = ausgeschlossene Diagnose, V = Verdachtsdiagnose, Z = symptomloser Zustand nach der betreffenden Diagnose). Auch die Lokalisation der Diagnose bei paarigen Organen oder Körperteilen wird kodiert: L = links, R = rechts, B = beidseitig. Des Weiteren sind im Abrechnungsfall die erbrachten Leistungen in Form eines Schlüsselcodes, dem Einheitlichen Bewertungsmaßstab (EBM), gespeichert. Für die Leistungen werden auch das Datum der Leistungsinanspruchnahme und die entstandenen Kosten (als Geld- und/oder Punktebetrag; s. Kapitel 1.7.3) gespeichert. Zusätzlich wird noch die Art der Inanspruchnahme für einen Abrechnungsfall gespeichert, also ob es sich um einen Original-, Sekundär-, Not- oder Vertretungsfall handelt. Für operative Eingriffe von ambulant tätigen Ärzten/Ärztinnen muss zusätzlich ein Operationen- und Prozedurenschlüssel (OPS, s. Kapitel 1.7.2) angegeben werden.

Einen genaueren Überblick, welche Daten im Rahmen von ambulanten Behandlungen für Forschungszwecke zur Verfügung stehen sollten, bietet Tabelle 1-4.

Tabelle 1-4. Daten zur ambulanten ärztlichen Versorgung (SGB V § 295) – Struktur und Merkmale

Merkmal	Erläuterung
Fall-Tabelle (ein Eintrag pro Abrechnungsfall)	
Versicherten-ID*	Identifikationsnummer des Versicherten (pseudonymisiert)
Fallnummer*	Identifikationsnummer eines Abrechnungsfalls
Berichtsjahr	Angabe des Jahres der Daten
Berichtsquartal	Angabe des Quartals der Daten
BSNR	Betriebsstättennummer des behandelnden Arztes (pseudonymisiert)
BSNR Überweiser	Ggf. Betriebsstättennummer des überweisenden Arztes (pseudonymisiert)
LANR Überweiser	Ggf. lebenslange Arztnummer des überweisenden Arztes (pseudonymisiert)
Beginn Fall	Beginn des Abrechnungsfalls (erstes Datum)
Ende Fall	Ende des Abrechnungsfalls (letztes Datum)
Art der Inanspruchnahme	O = Originalschein, V = Vertreterschein, N = Notfallschein, Z = Auftragsleistung, K = Konsiliaruntersuchung, M = Mit-/Weiterbehandlung
Unfallkennzeichen	0 = default, 2 = Unfall/-folgen, 3 = Versorgungsleiden
Fallwert Punktzahl	Summe der Punktwerte erbrachten Leistungen im Behandlungsfall

Merkmal	Erläuterung
Fallwert Kosten	Summe der Kosten im Behandlungsfall (in Euro)
EBM-Tabelle (Abrechnungsziffern, einer bis viele Einträge pro Behandlungsfall)	
Versicherten-ID*	Identifikationsnummer des Versicherten (pseudonymisiert)
Fallnummer*	Identifikationsnummer eines Abrechnungsfalls
LANR	Lebenslange Arztnummer des behandelnden Arztes (pseudonymisiert)
Gebührenordnungsnummer(n)	Gebührenordnungsnummer (gemäß EBM oder anderen Abrechnungskatalogen)
Datum der Leistung(en)	
Anzahl der Leistung(en)	
ICD-Tabelle (Diagnosen, einer bis viele Einträge pro Behandlungsfall)	
Versicherten-ID*	Identifikationsnummer des Versicherten (pseudonymisiert)
Fallnummer*	Identifikationsnummer eines Abrechnungsfalls
Diagnose	Nach aktuellem ICD-10-GM
Diagnosesicherheit	A = Ausschluss, G = Gesichert, V = Verdacht, Z = symptomloser Zustand nach Diagnose
Seitenlokalisation Diagnose	L = links, R = rechts, B = beidseitig
OPS-Tabelle (Operationen und Prozeduren, keiner bis viele Einträge pro Behandlungsfall)	
Versicherten-ID*	Identifikationsnummer des Versicherten (pseudonymisiert)
Fallnummer*	Identifikationsnummer eines Abrechnungsfalls
OPS-Code	Nach aktuellem OPS
Seitenlokalisation OPS	L = links, R = rechts, B = beidseitig

*** blau unterlegt**: Verknüpfungsrelevante Merkmale

1.2.2 Datenvolumen, Übermittlung, Verfügbarkeit sowie Nutzungshinweise

Das Potenzial von Daten zur ambulanten Versorgung liegt u.a. im Umfang der Daten. So haben, zumindest kurzzeitig, mehr als 90 % der Versicherten Kontakt zur ambulanten ärztlichen Versorgung (Grobe & Dräther, 2014). Informationen über Übermittlung, Umfang und Verfügbarkeit der Daten in den verschiedenen Datentabellen sind in Tabelle 1-5 zusammengefasst.

Tabelle 1-5. Daten zur ambulanten ärztlichen Versorgung (SGB V § 295) – Übermittlung und Verfügbarkeit der Daten

Datentabelle	Erläuterungen
Daten zur ambulanten ärztlichen Versorgung werden primär von niedergelassenen Ärzten/Ärztinnen dokumentiert, quartalsweise mit einer der 17 KVen abgerechnet und anschließend von den KVen an Krankenkassen übermittelt, wobei mit einem Zeitverzug von knapp 8 Monaten ab Quartalsende zu rechnen ist; bei mehr als 90 % der Bevölkerung wird pro Jahr mindestens eine Leistung abgerechnet	
Fall-Tabelle	Pro Abrechnungsfall ist genau ein Tabelleneintrag mit grundlegenden fallbezogenen Informationen vorhanden, wobei ein Fall alle Behandlungen in einer Arztpraxis innerhalb eines Quartals bei einem Patienten zusammenfassen; pro Person werden jährlich im Durchschnitt mehr als 8 Fälle abgerechnet
EBM-Tabelle	Pro Fall ist mindestens eine Abrechnungsziffer, i.d.R. gemäß EBM, im Sinne einer erbrachten Leistung dokumentiert; pro Person werden jährlich im Durchschnitt mehr als 60 Ziffern abgerechnet
ICD-Tabelle	Pro Fall ist mindestens eine Diagnose dokumentiert (seit 2000 gemäß jeweils gültiger ICD-10-Klassifikation); pro Person werden jährlich im Durchschnitt mehr als 35 Diagnoseeinträge mit den Daten übermittelt
OPS-Tabelle	Lediglich bei Abrechnung bestimmter Abrechnungsziffern müssen OPS-Codes zu durchgeführten Operationen im Sinne ergänzender Informationen zur Legitimation der EBM-Abrechnung dokumentiert werden, was etwa 1 % aller abgerechneten Fälle betrifft

Fehlende Daten

Bei der Nutzung der Daten ist zu beachten, dass nicht alle ambulanten Leistungen, die die Versicherten in Anspruch nehmen, auch in Daten der gesetzlichen Krankenkassen auftauchen. So gibt es ambulante Leistungen, die von anderen Stellen als den Krankenkassen finanziert werden, z. B private (Zusatz-)Versicherungen, aber auch der Bundesagentur für Arbeit (nach SGB II und SGB III), der gesetzlichen Rentenversicherung (nach SGB IV), der gesetzlichen Unfallversicherung (nach SGB VII) oder auch von den Versicherten selbst in Form individueller Gesundheitsleistungen (iGeL). Darüber hinaus existieren z.T. Selektivverträge zwischen Krankenkassen und bestimmten Ärzten/Ärztinnen, deren Abrechnungsdaten anderen Wegen folgen und nicht immer routinemäßig verfügbar sind.

1.3 Krankenhausbehandlungen

Daten aus Krankenhausbehandlungen enthalten Informationen zum Krankenhausaufenthalt, den ermittelten Diagnosen, durchgeführten Operationen und Prozeduren und den berechneten Entgelten. Die Behandlung kann voll- oder teilstationär erfolgen. Daten zu ambulanten Behandlungen in Krankenhäusern werden gesondert erfasst.

In Krankenhäusern werden Menschen zur Erkennung und Behandlung von Erkrankungen und zur Geburtshilfe behandelt und sind somit von Rehabilitations- und Vorsorgeeinrichtungen zu unterschieden (Gesundheitsberichtserstattung des Bundes, 2022a). In der Regel werden die Kosten für die Krankenhausbehandlungen bei gesetzlich Versicherten von den Krankenkassen übernommen.

In den Krankenhäusern finden vorwiegend stationäre Behandlungen statt, bei denen Patienten eine oder mehrere Nächte im Krankenhaus bleiben. Darüber hinaus gibt es auch teilstationäre Behandlungen, bei denen Patienten nur tagsüber oder nur nachts behandelt werden und danach in das häusliche Umfeld entlassen werden. Außerdem kommt es teilweise zu vor- oder nachstationären Behandlungen, also Besuche von Patienten im Krankenhaus zur Vor- oder Nachbereitung eines stationären Aufenthaltes, beispielsweise um die Behandlungsbedürftigkeit zu überprüfen oder den Behandlungserfolg zu sichern. In allen Fällen, in denen Leistungen im Krankenhaus zu Lasten der gesetzlichen Krankenversicherung durchgeführt werden, müssen die Krankenhäuser den Krankenkassen entsprechende Daten melden. Dies ist in §301 SGB V geregelt. Krankenhäuser können außerdem auf der Basis unterschiedlicher gesetzlicher Grundlagen auch ambulante Behandlungen durchführen. Daten hierzu werden typischerweise gesondert erfasst und an dieser Stelle nicht weiter erläutert.

DRG-System

Seit 2004 werden die stationären Leistungen im somatischen Bereich über ein Fallpauschalensystem abgerechnet, den sogenannten Diagnosis Related Groups (DRG). Dabei wird der stationäre Aufenthalt der Patienten anhand verschiedener Charakteristika einer Fallgruppe zugeordnet, für die eine pauschale Vergütung (berechnet auf Basis der Vorjahre) erstattet wird. Zu den Charakteristika zählen vor allem die Hauptdiagnose, der Schweregrad der Erkrankung, sowie das Alter der Patienten, das Gewicht (bei Neugeborenen) und weitere Charakteristika wie z. B Komplikationen. Um eine vergleichbare Dokumentation der entgeltrelevanten Daten zwischen den Krankenhäusern sicherzustellen (und somit eine gleiche Vergütung von vergleichbaren Krankenhausfällen), gibt es spezielle Kodierrichtlinien, um Diagnosen und Prozeduren möglichst einheitlich zu klassifizieren (Institut für das Entgeltsystem im Krankenhaus, 2022). In der Psychiatrie und Psychosomatik wurde in den letzten Jahren auch ein leistungsorientiertes, pauschalisierendes Vergütungssystem eingeführt, das pauschalisierende Entgeltsystem für Psychiatrie und Psychosomatik (PEPP). Dieses System dient allerdings nur der Budgetfindung, die Finanzierung psychiatrischer und psychosomatischer Krankenhausleistungen erfolgt weiterhin krankenhausindividuell.

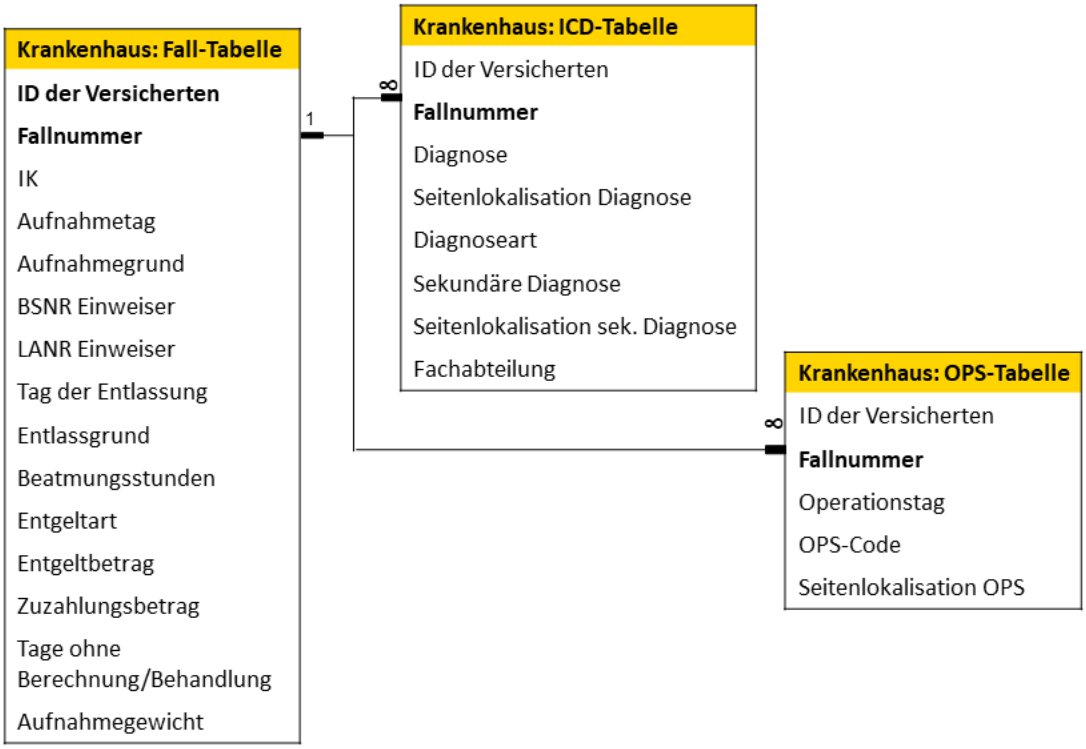

Abbildung 1-3. Schematische Darstellung der stationären Datentabellen

1.3.1 Struktur und Merkmalsumfang der Daten

Einen Überblick über die Datenstruktur der stationären Datentabellen bietet Abbildung 1-3. In den bei den Krankenkassen gespeicherten Daten aus Krankenhausbehandlungen finden sich grundlegende Informationen zum Krankenhausaufenthalt (Beginn/Ende, Dauer, Diagnosen). Des Weiteren sind Informationen zu den im Krankenhaus durchgeführten Operationen und Prozeduren gespeichert, die nach dem Operationen- und Prozedurenschlüssel kodiert sind (OPS, s.1.7.2). Für jede Operation oder Prozedur ist außerdem das Datum gespeichert und ggf. eine Seitenlokalisation (Grobe, Nimptsch, & Friedrich, 2014; Neubauer et al., 2017).

Für die Abrechnung werden außerdem die Entgeltschlüssel gespeichert. Ähnlich wie im ambulanten Bereich werden die erbrachten Leistungen also mithilfe eines Schlüsselsystems kodiert, wobei im stationären Bereich unterschiedliche Entgeltarten unterschieden werden können, z. B DRG-Fallpauschalen, aber auch Tagessätze oder Sonderentgelte. Darüber hinaus werden auch die daraus ermittelten Beträge und ggf. Zuzahlungen durch die Patienten gespeichert.

Behandlungsfall

Die zentrale Beobachtungseinheit bildet, ähnlich wie bei der ambulanten Versorgung, ein Behandlungsfall. Ein Behandlungsfall im stationären Bereich umfasst i.d.R. die Zeit zwischen Aufnahme und Entlassung der Patienten aus dem Krankenhaus. Allerdings müssen teilweise mehrere Aufenthalte einer Person in einem Krankenhaus zu einem Fall zusammengefasst werden, z. B. wenn die Patienten zwischenzeitlich in ein anderes Krankenhaus verlegt wurden. Auch bei teilstationären Aufenthalten werden mehrere Aufenthalte zu einem Fall zusammengefasst. Hier erfolgt i.d.R. eine quartalsweise Zählung wie im ambulanten Bereich. Eine genaue Beschreibung, wann und unter welchen Umständen

Krankenhausfälle zusammengeführt werden müssen, findet sich in den aktuellen Abrechnungsbestimmungen zum Fallpauschalensystem gemäß DRG[2] für Krankenhäuser beim Institut für das Entgeltsystem im Krankenhaus.

Aufnahme- und Entlassmeldung

Für jeden Behandlungsfall im Krankenhaus werden bei der Aufnahme der Patienten Daten an die Krankenkasse übermittelt, die Informationen zu behandelnder Einrichtung (über das Institutionskennzeichen (IK), s. Kapitel 1.8) und Fachabteilung, Aufnahmedatum, Aufnahmegrund, sowie ggf. zur einweisenden Institution bzw. Person enthalten. Der Aufnahmegrund erlaubt hierbei auch eine Unterscheidung zwischen voll- und teilstationären Aufenthalten. Bei der Entlassung wird das Entlassdatum, der Grund für die Entlassung und ggf. das Institutionskennzeichen der nachfolgend aufnehmenden Institution übermittelt. Darüber hinaus kann eine Beatmungsdauer (in Stunden) angegeben werden und eine unverbindliche Einstufung, ob die Patienten arbeitsfähig entlassen wurden. Eine Entlassmeldung wird vom Krankenhaus auch gemacht, wenn die Patienten verlegt werden. Somit kann es pro Behandlungsfall mehrere Entlassmeldungen geben. In der Regel existiert bei den Krankenkassen eine grundlegende, fallbezogene Tabelle, in der die letztendliche Entlassung mit Datum, Diagnosen und Entlassgrund aufgeführt ist (Grobe et al., 2014).

Diagnosen

Bei den Diagnosen lässt sich zwischen der Hauptdiagnose, die hauptsächlich für den stationären Aufenthalt verantwortlich ist, und Nebendiagnosen, die neben der Hauptdiagnose bestehen oder sich während des Aufenthalts entwickeln, unterscheiden (Gesundheitsberichtserstattung des Bundes, 2022b). Wenn es mehrere Nebendiagnosen gibt, stehen diese in den verfügbaren Tabellen alle gleichwertig nebeneinander, so dass sich keine Hierarchie ableiten lässt. Zusätzlich werden noch Einweisungs-, Aufnahmediagnosen unterschieden. Einweisungsdiagnosen können bei einer Einweisung ins Krankenhaus aus der ambulanten Versorgung mitgeteilt sein und sind nur optional vorhanden. Gleichfalls optional können nach ersten Einschätzungen im Krankenhaus Aufnahmediagnosen gestellt werden.

Einen genaueren Überblick, welche Daten im Rahmen von Krankenhausbehandlungen für Forschungszwecke zur Verfügung stehen sollten, bietet Tabelle 1-6.

Tabelle 1-6. Daten zur stationären Behandlung (SGB V §301) – Struktur und Merkmale

Merkmal	Erläuterung
Fall-Tabelle (ein Eintrag pro Krankenhausfall)	
Versicherten-ID*	Identifikationsnummer des Versicherten (pseudonymisiert)
Fallnummer*	Identifikationsnummer eines Abrechnungsfalls
IK	Institutionskennzeichen des behandelnden Krankenhauses (pseudonymisiert)
Aufnahmetag	Datum
Aufnahmegrund	Numerischer Schlüssel, der u.a. Unterscheidung in voll-/teilstationär und Normalfall/Unfall ermöglicht (verschlüsselt nach Technischer Anlage 2 Schlüssel 1[3])
BSNR Einweiser	Ggf. Betriebsstättennummer des einweisenden Arztes (pseudonymisiert)
LANR Einweiser	Ggf. lebenslange Arztnummer des einweisenden Arztes (pseudonymisiert)
Tag der Entlassung	Datum

2 abrufbar unter: https://www.g-drg.de/ag-drg-system-2025/abrechnungsbestimmungen/fpv-2025/fpv-2025, Stand: 04/2025
3 In den Technischen Anlagen (TA) der Richtlinien des Datenaustauschverfahrens zwischen Leistungserbringern und gesetzlicher Krankenversicherung sind die Schlüsselverzeichnisse einsehbar, s. https://www.gkv-datenaustausch.de/ (Stand 04/2025).

Merkmal	Erläuterung
Entlassgrund	Numerischer Schlüssel, der u.a. Identifikation von regulärer Entlassung, Versterben oder Verlegung ermöglicht (verschlüsselt nach Technischer Anlage 2 Schlüssel 1[1])
Beatmungsstunden	Ggf. Angabe zu Anzahl der Stunden mit Beatmung
Entgeltart	Art des Entgelts, z. B. DRG-Fallpauschale, Pflegesatz, etc.
Entgeltbetrag	Einzelbetrag zur Entgeltart
Zuzahlungsbetrag	Zuzahlungen durch den Patienten
Tage ohne Berechnung/ Behandlung	Bei Unterbrechung der Behandlung, z. B. Beurlaubung
Aufnahmegewicht	Angabe nur im ersten Lebensjahr
ICD-Tabelle (Diagnosen, min. ein Eintrag bis viele Einträge pro Krankenhausfall)	
Versicherten-ID*	Identifikationsnummer des Versicherten (pseudonymisiert)
Fallnummer*	Identifikationsnummer eines Abrechnungsfalls
Diagnose	Diagnose nach ICD-10-GM
Seitenlokalisation Diagnose	L = links, R = rechts, B = beidseitig
Diagnoseart	Art der ICD-Angabe: Einweisungsdiagnose, Aufnahmediagnose, Hauptdiagnose, Entlass-/Verlegungsdiagnose, Nebendiagnose
Sekundäre Diagnose	Diagnose nach ICD-10-GM
Seitenlokalisation sekundäre Diagnose	L = links, R = rechts, B = beidseitig
Fachabteilung	4-stelliger Code, der Identifikation der Fachabteilung ermöglicht
OPS-Tabelle (Operationen und Prozeduren, keiner bis viele Einträge pro Krankenhausfall möglich)	
Versicherten-ID*	Identifikationsnummer des Versicherten (pseudonymisiert)
Fallnummer*	Identifikationsnummer eines Abrechnungsfalls
Operationstag	Datum der Operation/Prozedur
OPS-Code	Nach aktuellem OPS
Seitenlokalisation OPS	L = links, R = rechts, B = beidseitig

* **blau unterlegt**: Verknüpfungsrelevante Merkmale

1.3.2 Datenvolumen, Übermittlung, Verfügbarkeit sowie Nutzungshinweise

Umfang und Validität der Daten

Die Daten aus dem stationären Bereich, die bei den Krankenkassen zu Abrechnungszwecken vorliegen, ermöglichen z. B. Aussagen über den Outcome von Krankenhausbehandlungen oder die Analyse von Behandlungskosten (z. B. Motzek, Werblow, Schmitt, & Marquardt, 2019; Swart, Deh, & Robra, 2008). Die Daten zu einem Behandlungsfall können u.U. sehr umfangreich sein, da eine Entlassmeldung auch bei einer Verlegung gemacht werden muss. Gerade bei langen Krankenhausaufenthalten mit mehrfachen (Rück-)Verlegungen sind für einen Behandlungsfall dann viele Informationen und eine Vielzahl an Haupt- und Nebendiagnosen gespeichert sind. Dies ermöglicht aber auch, die komplexen Verlegungshistorien nachzuvollziehen (Grobe et al., 2014). Informationen über Übermittlung, Umfang und Verfügbarkeit der Daten in den verschiedenen Datentabellen sind in Tabelle 1-7 zusammengefasst.

Tabelle 1-7. Daten zur stationären Behandlung (SGB V §301) – Übermittlung und Verfügbarkeit

Datentabelle	Erläuterungen
Von den Krankenhäusern an die Krankenkassen gemeldet; eine Meldung muss u.a. bei Aufnahme, Verlegung und Entlassung gemacht werden; es ist mit einem Zeitverzug von etwa 3 Monaten nach der Entlassung zu rechnen, bis die Daten bei den Krankenkassen vollständig verfügbar sind	
Fall-Tabelle	pro Fall ist genau ein Tabelleneintrag mit grundlegenden fallbezogenen Informationen vorhanden, wobei ein Fall i.d.R. den Zeitraum zwischen Aufnahme und Entlassung aus dem Krankenhaus umfasst; 2023 wurden allein im BKK-System 1,7 Mio. voll- oder teilstationäre Krankenhausaufenthalte abgerechnet. Das entspricht etwa 17,6 Fällen pro 100 Versicherten (BKK Gesundheitsreport 2024, S. 189).
ICD-Tabelle	für jeden Fall muss mindestens eine Behandlungshauptdiagnose dokumentiert werden (seit 2000 gemäß jeweils gültiger ICD-10-Klassifikation); ggf. werden zusätzlich Nebendiagnosen dokumentiert, sowie die Diagnosen, die bei Einweisung und Aufnahme vergeben wurden.
OPS-Tabelle	Dokumentation der Operationen und Prozeduren inklusive tagesgenauem Datum, sofern während des Krankenhausaufenthalts durchgeführt

Bei der Auswertung der Diagnosen in Krankenhausbehandlungen spielt vor allem die Entlassungsdiagnosen eine Rolle, wobei die fallbezogen dokumentierte Hauptdiagnose den maßgeblichsten Anlass des Krankenhausaufenthalts kennzeichnen sollte. Einweisungs- und Aufnahmediagnosen können als weniger valide angesehen werden und sich u.U. deutlich von abschließend dokumentierten Diagnosen unterscheiden (Grobe et al., 2014; Neubauer et al., 2017).

Dauer des Krankenhausaufenthalts

Da mitunter mehrere Krankenhausaufenthalte zu einem Behandlungsfall zusammengefasst werden müssen (s.o.), lässt sich die Aufenthaltszeit der Patienten im Krankenhaus nicht immer über die Differenz zwischen Aufnahme- und Entlassdatum berechnen. Es gibt jedoch unterschiedliche Möglichkeiten, Zeiten zwischen Teilaufenthalten oder Behandlungspausen angemessen zu berücksichtigen (s. Grobe et al., 2014).

1.4 Arzneimittel

Für Arzneimittel, die Versicherte in Apotheken erhalten, werden abrechnungsrelevante Daten an die Krankenkassen an die Krankenkassen gemeldet, z. B. welches Arzneimittel wann verordnet und ausgegeben wurde. Dies gilt allerdings i.d.R. nur für verschreibungspflichtige Medikamente, da die Kosten für frei verkäufliche Arzneimittel i.d.R. von den Versicherten selbst getragen werden müssen.

Arzneimittelgruppen

Bei den Arzneimitteln lassen sich grundsätzlich drei Gruppen unterscheiden. Zunächst gibt es die OTC-Arzneimittel (OTC = over the counter), wozu alle nicht-apothekenpflichtigen Arzneimittel (die auch im Supermarkt oder der Drogerie verkauft werden dürfen) sowie auch bestimmte apothekenpflichtige Arzneimittel gehören, die ohne Rezept erworben werden können. Als zweite Gruppe gibt es die verschreibungspflichtigen Arzneimittel, die in Apotheken nur bei Vorlage einer ärztlichen Verordnung erhältlich sind (§43ff. Arzneimittelgesetz). Zuletzt gibt es noch die Betäubungsmittel, die nur in der Apotheke erhältlich sind und bei denen eine besondere Verordnung notwendig ist (§13 Betäubungsmittelgesetz).

Bei Arzneimitteln gibt es zu einem Wirkstoff häufig verschiedene Präparate, z. B. von unterschiedlichen Herstellern oder in unterschiedlichen Formen oder Größen. Jedes dieser unterschiedlichen Präparate aus dem pharmazeutischen Bereich ist dabei eindeutig durch eine Pharmazentralnummer (PZN) gekennzeichnet, die zu Abrechnungszwecken auch an die Krankenkasse gemeldet wird. Den PZN lassen sich dann über umfangreiche Hilfstabellen wieder der genaue Präparatename, Hersteller, Wirkstoffe und Wirkstoffgruppenzuordnungen sowie Packungsgrößen herleiten (s. Kapitel 1.7.4). Die PZN ist damit eine zentrale Information.

Organisation der Abrechnung bei den Apotheken

Daten zu den in Apotheken abgegebenen Arzneimitteln landen i.d.R. nur bei der Krankenkasse, wenn diese auch die Kosten für die Arzneimittel trägt. Dies ist normalerweise nur bei verschreibungspflichtigen Medikamenten der Fall, wobei es auch Ausnahmen für nicht-verschreibungspflichtige Medikamente gibt (s. Arzneimittelrichtlinie des GB-A, (Banz AT 05.07.2022 B1)). Wenn die Kosten für ein Medikament von den Krankenkassen übernommen werden, erhalten die Versicherten in der Arztpraxis ein Rezeptblatt, das sie in der Apotheke vorlegen, um das Arzneimittel zu erhalten. Die Apotheken leiten die Daten, die auf dem Rezeptblatt abgedruckt sind, dann i.d.R. über zwischengeschaltete Rechenzentren an die Krankenkassen weiter, um die Kosten erstattet zu bekommen (Schröder, 2014). Die Informationsweitergabe ist in §300 SGB V geregelt.

1.4.1 Struktur und Merkmalsumfang der Daten

Die Krankenkassen erhalten von den Apotheken Informationen zum abgegebenen Arzneimittel in Form der jeweiligen PZN, das Abgabedatum, sowie Informationen zum verordnenden Arzt/zur verordnenden Ärztin und dem Verordnungsdatum. Außerdem werden die Preise der Arzneimittel gespeichert, wobei hier zwischen Bruttopreisen (entspricht dem Apothekenabgabepreis) und den für die Krankenkasse tatsächlich anfallenden Nettopreisen unterschieden werden kann. Bei den Nettopreisen werden Zuzahlungen der Versicherten und eventuelle Rabatte abgezogen. Auch die Zuzahlungen, die die Versicherten leisten müssen, werden i.d.R. ausgewiesen (Neubauer et al., 2017; Schröder, 2014).

Einen genaueren Überblick, welche Daten für Arzneimittel zu Forschungszwecken zur Verfügung stehen sollten, bietet Tabelle 1-8.

Tabelle 1-8. Arzneimitteldaten (SGB V §300) – Struktur und Merkmale

Merkmal	Erläuterung
(kein, ein oder mehrere Einträge pro Versicherten-ID möglich)	
Versicherten-ID	Identifikationsnummer des Versicherten (pseudonymisiert)
Apothekennummer / IK	Institutionskennzeichen/Nummer der Apotheke (pseudonymisiert)
BSNR des verordnenden Arztes	Betriebsstättennummer des verordnenden Arztes (pseudonymisiert)
LANR des verordnenden Arztes	Lebenslange Arztnummer des verordnenden Arztes (pseudonymisiert)
PZN	Pharmazentralnummer
ATC-Code	Amtlicher Code gemäß ATC-Klassifikationssystem
DDD	Daily defined dose
Datum der Ausstellung	Datum, an dem das Arzneimittel verordnet wurde
Datum der Abgabe	Datum, an dem das Arzneimittel ausgegeben wurde
Kosten	Kosten des Arzneimittels
Zuzahlung	Höhe der Zuzahlung des Versicherten

1.4.2 Datenvolumen, Übermittlung, Verfügbarkeit sowie Nutzungshinweise

Mithilfe der Arzneimitteldaten der Krankenkassen können z. B. das Verordnungsverhalten und die Kosten, die für Arzneimittel anfallen, analysiert werden. Informationen zur Übermittlung, Umfang und Verfügbarkeit der Daten sind in Tabelle 1-9 zusammengefasst.

Tabelle 1-9. Arzneimitteldaten (SGB V §300) – Übermittlung und Verfügbarkeit

Datentabelle	Erläuterungen
Arzneimittel-Daten	Von den Apotheken an die Krankenkassen übermittelt, ungeprüft sind die Daten ca. 2 Monate später verfügbar; geprüft ist mit einem Datenverzug von bis zu 1 Jahr zu rechnen; 72,7 % der BKK-Versicherten wurde im Jahr 2023 mindestens ein Arzneimittel verordnet. Jeder Versicherte erhielt im Durchschnitt 7,9 Arzneimittelverordnungen bzw. 546 DDD (BKK Gesundheitsreport 2024, S. 253).

Fehlende Daten

In den Arzneimitteldaten der Krankenkassen sind nur solche Arzneimittel enthalten, die auch über die Krankenkassen abgerechnet werden. Arzneimittel, die die Patienten selbst zahlen, tauchen in den Daten nicht auf. Das gilt vor allem für die sogenannte OTC-Präparate, die nicht-verschreibungspflichtig sind und deren Kosten nur in Ausnahmefällen von Krankenkassen übernommen werden. Auch Arzneimittel, die in Krankenhäusern abgegeben werden, sind nicht enthalten. Hier besteht lediglich die Möglichkeit, hochpreisige Arzneimittel über den OPS zu kodieren und so Informationen über die ausgegebenen Arzneimittel zu erhalten (Neubauer et al., 2017).

1.5 Heilmittel

Unter Heilmitteln versteht man persönlich erbrachte, medizinische Leistungen wie Physio-, Ergo- oder Logopädie. Zu Abrechnungszwecken werden Daten von den Leistungserbringern an die Krankenkassen übermittelt, z. B. zur Art und Umfang der Leistung, dem Tag der Leistungserbringung und zur Indikation.

Heilmittel sind persönlich erbrachte, medizinische Leistungen wie Physikalische Therapie, Ergotherapie, Stimm-, Sprech-, Sprach- und Schlucktherapie, Ernährungstherapie oder Podologie. Im Heilmittelkatalog (nach §125 SGB V) ist festgelegt, welche Heilmittel bei welcher Indikation verordnet werden können und somit auch von den Krankenkassen erstattet werden. Der Heilmittelkatalog[4] kann beim Gemeinamen Bundesausschuss (G-BA) eingesehen werden. Damit die Kosten von den Krankenkassen übernommen werden, ist immer eine ärztliche Verordnung des Heilmittels notwendig. Die Leistungserbringer, also z. B. Ergotherapeuten/Ergotherapeutinnen oder Physiotherapeuten/Physiotherapeutinnen, müssen entsprechend ausgebildet sein und benötigen eine Zulassung der Krankenkassen, damit sie die Leistungen abrechnen können. Die Abrechnung der Heilmittel fällt unter §302 SGB V.

1.5.1 Struktur und Merkmalsumfang der Daten

Den Krankenkassen werden Daten zur Art der Leistung und dem Tag der Leistungserbringung gemeldet. Die Art der durchgeführten Therapie wird mit einer fünfstelligen Positionsnummer codiert, deren Aufbau in den Technischen Anlagen der Richtlinien des Datenaustauschverfahrens zwischen Leistungserbringern und gesetzlicher Krankenversicherung geregelt ist. Darüber hinaus erhalten die Krankenkassen Daten zum Preis der Leistung, sowie Informationen über Zuzahlungen der Versicherten. Bei Heilmitteln müssen die Versicherten i.d.R. eine Zuzahlung von 10 % der Therapiekosten und pauschal 10 Euro je Verordnung zahlen. Außerdem gibt es Informationen zur Verordnung, d. h. zum Datum der Verordnung und zu den verordnenden Ärztinnen und Ärzten. Diese Informationen sind auf dem Verordnungsblatt gedruckt, das die Versicherten vom verordnenden Arzt/von der verordnenden Ärztin erhalten und werden von Leistungserbringern auf elektronischem Weg an die Krankenkassen gemeldet. Zusätzlich wird für Heilmittel i.d.R. auch ein Indikationsschlüssel aus dem Heilmittelkatalog angegeben, der den medizinischen Grund für das Heilmittel kodiert. Der Schlüssel entspricht nicht dem ICD-Code für Diagnosen, sondern ist eigenständig. Der Code „EX1a“ bedeutet z. B. Verletzung/Operation/Erkrankung der Extremitäten (EX) mit prognostisch kurzzeitigem Behandlungsbedarf (1) und der Leitsymptomatik Bewegungsstörungen.

Einen genaueren Überblick, welche Daten zu Heilmitteln für Forschungszwecke zur Verfügung stehen sollten, bietet Tabelle 1-10.

4 abrufbar unter: https://www.g-ba.de/richtlinien/12/, Stand: 01/2025

Tabelle 1-10. Heilmitteldaten (SGB V §302) – Struktur und Merkmale

Merkmal	Erläuterung
(kein, ein oder mehrere Einträge pro Versicherten-ID möglich)	
Versicherten-ID	Identifikationsnummer des Versicherten (pseudonymisiert)
BSNR des verordnenden Arztes	Betriebsstättennummer des verordnenden Arztes (pseudonymisiert)
LANR des verordnenden Arztes	Lebenslange Arztnummer des verordnenden Arztes (pseudonymisiert)
Datum der Verordnung	Datum, an dem das Heilmittel verordnet wurde
Indikationsschlüssel	Code für die Indikation des Heilmittels laut Heilmittelkatalog
Datum der Erbringung	Datum der Leistungsbringung
Positionsnummer	Art des Heilmittels
Kosten	Kosten
Zuzahlung	Höhe der Zuzahlung des Versicherten

1.5.2 Datenvolumen, Übermittlung, Verfügbarkeit sowie Nutzungshinweise

Aus den Daten zu Heilmitteln lässt sich analysieren, welche Therapien verordnet wurden und welche Kosten verursacht wurden, aber z. B. auch, welche Facharztgruppen die Leistungen verordnen. Diese Analysen werden teilweise auch von den Krankenkassen selbst an den eigenen Daten durchgeführt, wie z. B. von der Barmer, die jährlich einen Heil- bzw. Hilfsmittelreport veröffentlicht (Schmitt & Wende, 2021, 2022). Auch hier ist allerdings zu beachten, dass Heilmittel auch über andere Kostenträger abgerechnet werden können (z. B. über die Berufsgenossenschaften) und dann nicht in den Daten der Krankenkassen auftauchen. Informationen über Übermittlung, Umfang und Verfügbarkeit der Daten sind in Tabelle 1-11 zusammengefasst.

Tabelle 1-11. Heilmitteldaten (SGB V §302) – Übermittlung und Verfügbarkeit

Datentabelle	Erläuterungen
Heilmittel-Daten	Von den Leistungserbringern an die Krankenkassen übermittelt; es ist mit einem Datenverzug von ca. 7 Monaten nach Durchführung der Therapie zu rechnen; GKV-weit wurden 2023 38,9 Mio. Verordnungsblätter für Heilmittel ausgestellt (GKV-SV 2024, S. 16).

1.6 Hilfsmittel

Hilfsmittel sind technische oder andere medizinisch notwendige Produkte, die den Versicherten helfen sollen, die Krankenbehandlung zu sichern oder einer Behinderung vorzubeugen bzw. sie abzumildern. Die Kosten für notwendige Hilfsmittel werden von den Krankenkassen übernommen. Die meisten Produkte, für die die Kosten übernommen werden, sind im Hilfsmittelverzeichnis aufgeführt.

Unter den Begriff der Hilfsmittel fallen eine Vielzahl verschiedener Produkte, die die Krankenbehandlung sichern bzw. einer Behinderung vorbeugen oder sie ausgleichen. Dazu zählen z. B. Sehhilfen, Hörhilfen, orthopädische Hilfsmittel, Rollstühle, Kompressionsstrümpfe, Spritzen oder Inhalationsgeräte. Im Hilfsmittelverzeichnis [5]sind Produkte aufgeführt, für die die Kosten übernommen werden. Das Verzeichnis ist allerdings nicht abschließend, d. h. es können auch Kosten für andere Produkte übernommen werden, wenn dies von der Krankenkasse genehmigt wird. Die Hilfsmittel werden i.d.R. von Apotheken oder dem Sanitätsfachhandel ausgegeben. Die Abrechnung der Hilfsmittel fällt unter §302 SGB V.

1.6.1 Struktur und Merkmalsumfang der Daten

Den Krankenkassen werden Daten zur Art des Hilfsmittels, dem Tag der Leistungserbringung bzw. -bereitstellung und ggf. der Menge gemeldet. Die Art des Hilfsmittels wird mit einer zehnstelligen Positionsnummer codiert, die im Hilfsmittelkatalog nachgeschlagen werden kann. Das Verzeichnis ist nach Produktgruppen sortiert, die in Tabelle 1-13 aufgelistet sind. Für Hilfsmittel, die nicht im Hilfsmittelverzeichnis aufgeführt sind, gibt es Sonderpositionsnummern, über die ggf. zumindest die Produktart identifizierbar ist. Darüber hinaus erhalten die Krankenkassen Daten zum Preis der Leistung, sowie Informationen über Zuzahlungen der Versicherten. Außerdem gibt es Informationen zur Verordnung, d. h. zum Datum der Verordnung und zu den verordnenden Ärzten/Ärztinnen. Diese Informationen sind auf dem Verordnungsblatt abgedruckt, das die Versicherten vom verordnenden Arzt/von der verordnenden Ärztin erhalten und werden von Leistungserbringern auf elektronischem Weg an die Krankenkassen gemeldet.

Einen genaueren Überblick, welche Daten zu Hilfsmitteln für Forschungszwecke zur Verfügung stehen sollten, bietet Tabelle 1-12.

5 abrufbar unter: https://hilfsmittel.gkv-spitzenverband.de/home, Stand 01/2025

Tabelle 1-12. Hilfsmitteldaten (SGB V §302) – Struktur und Merkmale

Merkmal	Erläuterung
(kein, ein oder mehrere Einträge pro Versicherten-ID möglich)	
Versicherten-ID	Identifikationsnummer des Versicherten (pseudonymisiert)
BSNR des verordnenden Arztes	Betriebsstättennummer des verordnenden Arztes (pseudonymisiert)
LANR des verordnenden Arztes	Lebenslange Arztnummer des verordnenden Arztes (pseudonymisiert)
Datum der Verordnung	Datum, an dem das Hilfsmittel verordnet wurde
Datum der Abgabe / Erbringung	Datum, an dem das Hilfsmittel ausgegeben wurde
Positionsnummer	Art des Hilfsmittels
Kennzeichen Hilfsmittel	Zweistelliger Code, der z. B. Neulieferung, Reparatur, Wartung, Wiedereinsatz eines Hilfsmittels kennzeichnet
Kosten	Kosten
Zuzahlung	Höhe der Zuzahlung des Versicherten

Tabelle 1-13. Produktgruppen des Hilfsmittelkatalogs.

Produkt-gruppe	Bezeichnung		Produkt-gruppe	Bezeichnung
01	Absauggeräte		24	Beinprothesen
02	Adaptionshilfen		25	Sehhilfen
03	Applikationshilfen		26	Sitzhilfen
04	Bade- und Duschhilfen		27	Sprechhilfen
05	Bandagen		28	Stehhilfen
06	Bestrahlungsgeräte		29	Stomaartikel
07	Blindenhilfsmittel		30	NN NN
08	Einlagen		31	Schuhe
09	Elektrostimulationsgeräte		32	Therapeutische Bewegungsgeräte
10	Gehhilfen		33	Toilettenhilfen
11	Hilfsmittel gegen Dekubitus		34	Haarersatz
12	Hilfsmittel bei Tracheostoma und Laryngektomie		35	Epithesen
			36	Augenprothesen
13	Hörhilfen		37	Brustprothesen
14	Inhalations- und Atemtherapiegeräte		38	Armprothesen
15	Inkontinenzhilfen		50	Pflegehilfsmittel zur Erleichterung der Pflege
16	Kommunikationshilfen			
17	Hilfsmittel zur Kompressionstherapie		51	Pflegehilfsmittel zur Körperpflege/ Hygiene
18	Kranken-/Behindertenfahrzeuge			
19	Krankenpflegeartikel		52	Pflegehilfsmittel zur selbstständigeren Lebensführung/Mobilität
20	Lagerungshilfen			

21	Messgeräte für Körperzustände/ -funktionen		54	Zum Verbrauch bestimmte Pflegehilfs-mittel
22	Mobilitätshilfen		99	Verschiedenes
23	Orthesen/Schienen			

1.6.2 Datenvolumen, Übermittlung, Verfügbarkeit sowie Nutzungshinweise

Obwohl Unterschiede in der Leistung und Abrechnung von Heilmitteln und Hilfsmittel bestehen, werden die Daten von den Krankenkassen häufig in einem gemeinsamen Datawarehouse kombiniert. Informationen über Übermittlung, Umfang und Verfügbarkeit der Daten sind in Tabelle 1-14 zusammengefasst.

Tabelle 1-14. Hilfsmitteldaten (SGB V §302) – Übermittlung und Verfügbarkeit

Datentabelle	Erläuterungen
Hilfsmittel-Daten	Von den Leistungserbringern an die Krankenkassen übermittelt; es ist mit einem Datenverzug von ca. 7 Monaten nach Abgabe des Hilfsmittels zu rechnen; 2022 wurden ca. 18,1 Mio. GKV-Versicherte mit einem Hilfsmittel aus dem Hilfsmittelverzeichnis versorgt (bifg 2024, S. 10).

1.7 Klassifikationssysteme

Ein Großteil der Informationen in Routinedaten liegt i.d.R. nicht als Freitext vor, sondern in Form von numerischen oder alphanumerischen Schlüsselcodes, die mithilfe eines entsprechenden Klassifikationssystems übersetzt werden können. Die Klassifikationssysteme sind teilweise sehr umfangreich, da sie komplexe Sachverhalte (z. B. alle möglichen Erkrankungen und Verletzungen oder Behandlungen) in eindeutige Codes überführen müssen, d. h. in eine begrenzte Auswahl sich gegenseitig ausschließender Kategorien. Die Klassifikationssysteme ermöglichen eine einfache, schnelle und systematische Kommunikation. So ist z. B. mit der Kurzformel für eine Erkrankung nach ICD eine ganze Reihe an Informationen verbunden. Der Code „J10.0“ steht z. B. für Grippe mit Lungenentzündung durch saisonale Influenzaviren, womit ersichtlich ist, welche Leitsymptome vorliegen und was die Ursache der Erkrankung ist.

In den nachfolgenden Abschnitten werden die wichtigsten Klassifikationssysteme vorgestellt. Die Klassifikationssysteme werden regelmäßig überarbeitet, d. h. die Codes bzw. die Bedeutung der Codes kann sich von Jahr zu Jahr unterscheiden. Für jedes System wird auch ein Verweis zu einer Online-Ressource angegeben (Stand: 03/2025), wo der aktuelle Klassifikationskatalog abgerufen werden kann.

1.7.1 ICD

Einen wesentlichen Informationsbestandteil von Routinedaten zur gesundheitlichen Versorgung bilden Angaben zu Erkrankungsdiagnosen. Diese werden vorrangig im Rahmen der ambulanten ärztlichen Behandlung sowie bei Behandlungen in Krankenhäusern dokumentiert, um die Notwendigkeit von Behandlungen zu begründen und damit letztendlich auch Abrechnungen zu legitimieren. Um Erkrankungen einheitlich zu bezeichnen und zu codieren, dient in der ambulanten und stationären Versorgung in Deutschland aktuell die **Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme, 10. Revision, German Modification (ICD-10-GM)** (Stand 12/2024 in der Version 2024). In diesem Katalog, der jährlich überarbeitet wird, sind weitestgehend alle bekannten Krankheiten verzeichnet und mit einem Schlüsselcode versehen. Die im deutschen Gesundheitssystem verwendete Variante des ICD basiert auf dem von der WHO (World Health Organization) regelmäßig überarbeiteten Klassifikationssystem.

Der ICD ist hierarchisch aufgebaut und in 22 Krankheitskapitel gegliedert, die mit Buchstaben gekennzeichnet sind (z. B. J: Krankheiten der Atmungsorgane, S: Verletzungen und Vergiftungen; s. Tabelle 1-15). Innerhalb eines Kapitels werden die Diagnosen dann mithilfe von zwei Ziffern differenziert, so dass dreistellige ICD-Codes (ein Buchstabe + zwei Ziffern) Krankheiten auf allgemeiner Ebene unterscheiden (z. B. J45: Asthma bronchiale oder S01: offene Wunde des Kopfes). Mithilfe von ein oder zwei weiteren Ziffern lässt sich die Diagnose noch verfeinern (z. B.: J45.0: vorwiegend allergisches Asthma bronchiale, J45.1: nichtallergisches Asthma bronchiale oder S01.31: offene Wundes des Kopfes, genauer an der Ohrmuschel). Durch die zunehmende Differenzierung der Diagnosen auf den unteren Hierarchieebenen lässt sich eine große Anzahl an Krankheiten klassifizieren. So sind auf dreistelliger Ebene ca. 2.000 Krankheitsklassen verzeichnet und auf vierstelliger Ebene mehr als 12.000 Diagnosen.

Tabelle 1-15. Krankheitskapitel des ICD-10

Kapitel	Gliederung auf 3-stelliger Ebene	Titel
I	A00-B99	Bestimmte infektiöse und parasitäre Krankheiten
II	C00-D48	Neubildungen
III	D50-D90	Krankheiten des Blutes und der blutbildenden Organe, sowie bestimmte Störungen mit Beteiligung des Immunsystems
IV	E00-E90	Endokrine, Ernährungs- und Stoffwechselkrankheiten
V	F00-F99	Psychische und Verhaltensstörungen
VI	G00-G99	Krankheiten des Nervensystems
VII	H00-H59	Krankheiten des Auges und der Augenanhangsgebilde
VIII	H60-H95	Krankheiten des Ohres und des Warzenfortsatzes
IX	I00-I99	Krankheiten des Kreislaufsystems
X	J00-J99	Krankheiten des Atmungssystems
XI	K00-K93	Krankheiten des Verdauungssystems
XII	L00-L99	Krankheiten der Haut und der Unterhaut
XIII	M00-M99	Krankheiten des Muskel-Skelett-Systems und des Bindegewebes
XIV	N00-N99	Krankheiten des Urogenitalsystems
XV	O00-O99	Schwangerschaft, Geburt und Wochenbett
XVI	P00-P96	Bestimmte Zustände, die ihren Ursprung in der Perinatalperiode haben
XVII	Q00-Q99	Angeborene Fehlbildungen, Deformitäten und Chromosomenanomalien
XVIII	R00-R99	Symptome und abnorme klinische und Laborbefunde, die anderenorts nicht klassifiziert sind
XIX	S00-T98	Verletzungen, Vergiftungen und bestimmte andere Folgen äußerer Ursachen
XX	V01-Y84	Äußere Ursachen von Morbidität und Mortalität
XXI	Z00-Z99	Faktoren, die den Gesundheitszustand beeinflussen und zur Inanspruchnahme des Gesundheitswesens führen
XXII	U00-U85	Schlüsselnummern für besondere Zwecke

Neben den Krankheitskapiteln gibt es noch Kapitel, mit deren Hilfe weitere Besonderheiten kodiert werden können, die v.a. für Behandlungs- und Abrechnungszwecke relevant sind. Mithilfe von Kapitel XXI lassen sich z. B. Situationen kodieren, die den Gesundheitszustand beeinflussen und zur Inanspruchnahme des Gesundheitswesens führen, aber keine Erkrankungen sind, z. B. Z00.1: Gesundheitsvorsorgeuntersuchung eines Kindes oder Z03.4: Beobachtung bei Verdacht auf einen Herzinfarkt. Für besondere Fälle stehen die Schlüsselnummern mit dem Buchstaben U zur Verfügung, z. B. für die vorläufige Zuordnung von Krankheiten (z. B. U07.1!: Covid 19).

Auf den Seiten des Bundesinstituts für Arzneimittel und Medizinprodukte (BfArM) findet sich der aktuelle ICD-Klassifikationskatalog[6], der zurzeit jährlich überarbeitet wird.

6 abrufbar unter: https://www.bfarm.de/DE/Kodiersysteme/Klassifikationen/ICD/ICD-10-GM/_node.html, Stand 01/2025

1.7.2 OPS

Mithilfe des **Operationen- und Prozedurenschlüssels (OPS)** werden Operationen, Prozeduren und allgemein medizinischen Maßnahmen im stationären Bereich und im Bereich des ambulanten Operierens verschlüsselt. Im OPS-Katalog gibt es sechs Hauptkapitel, die nicht durchgehend nummeriert sind (s. Tabelle 1-16). Durch Anhängen weiterer Ziffern lassen sich die Hauptkategorien mehr und mehr verfeinern, z. B. 5-36: Operation an den Koronargefäßen, 5-360.1 Endarteriektomie der Koronararterien, offen chirurgisch, mit Patch. Der OPS-Katalog ist sehr umfangreich mit derzeit mehr als 33.000 Einträgen.

Auf den Seiten des Bundesinstituts für Arzneimittel und Medizinprodukte findet sich der aktuelle OPS-Katalog[7], der zurzeit jährlich überarbeitet wird.

Tabelle 1-16. Hauptkapitel des Operationen- und Prozedurenschlüssels (OPS)

Kapitel	Gliederung	Titel
1	1-10 … 1-99	Diagnostische Maßnahmen
3	3-03 … 3-99	Bildgebende Diagnostik
5	5-01 … 5-99	Operationen
6	6-00 … 6-00	Medikamente
8	8-01 … 8-99	Nicht operative therapeutische Maßnahmen
9	9-26 … 9-99	Ergänzende Maßnahmen (wie geburtsbegleitende oder psychotherapeutische Maßnahmen)

1.7.3 EBM

Mithilfe des **Einheitlichen Bewertungsmaßstab (EBM)** werden ambulante Leistungen von Vertragsärzten/-ärztinnen und Vertragspsychotherapeuten/-innen in Deutschland abgerechnet. Im EBM werden über 2.500 ambulante Einzelleistungen beschrieben, die sogenannten Gebührenordnungspositionen (GOP), die von den Leistungserbringern abgerechnet werden können. Zum Teil sind die Leistungen auch mit einer Angabe zum erforderlichen Zeitaufwand verstehen. Jede GOP ist fünfstellig und mit einem Punktwert versehen. Der Punktwert stellt das wertemäßige Verhältnis zwischen den GOPs dar. Grob vereinfacht wird der Punktwert dann mit einem (jährlich angepassten) Eurocent-Wert multipliziert, um die Vergütung für die Leistung zu erhalten.

Im EBM gibt es arztgruppenübergreifende Abschnitte und arztgruppenspezifischen (fachärztlichen) Abschnitte (s. Tabelle 1-17). Zusätzlich gibt es einen Abschnitt mit Kostenpauschalen für Sachkosten. Die einzelnen GOPs sind i.d.R. mit Bedingungen oder Beschränkungen versehen, d. h. nicht alle GOPs können zeitgleich nebeneinander abgerechnet werden. Teilweise gibt es auch Beschränkungen, wie häufig GOPs im Quartal abgerechnet werden können. So gibt es z. B. eine Grundpauschale für den Arzt-Patienten-Kontakt, die pro Patienten einmal im Quartal abgerechnet werden kann.

Der EBM-Katalog [8]wird fortlaufend überarbeitet. Dadurch können sich die Abrechnungsmodalitäten ggf. von Quartal zu Quartal verändern, z. B. wenn neue GOPs hinzukommen, sich Vergütungen ändern oder die Bedingungen und Beschränkungen der GOPs angepasst werden. Die für das aktuelle Quartal relevante Version und die vergangenen Versionen finden sich auf den Seiten der Kassenärztlichen Bundesvereinigung (KBV).

7 abrufbar unter: https://www.bfarm.de/DE/Kodiersysteme/Klassifikationen/OPS-ICHI/OPS/_node.html, Stand 01/2025
8 abrufbar unter: https://www.kbv.de/html/online-ebm.php, Stand 01/2025

Tabelle 1-17. Kapitel des Einheitlichen Bewertungsmaßstabs (EBM) und ausgewählte, beispielhafte Unterkapitel

Kapitel	Titel	Ausgewählte Unterkapitel
I	Allgemeine Bestimmungen	
II	Arztgruppenübergreifende allgemeine Gebührenordnungspositionen	1.1 Aufwandserstattung für die besondere Inanspruchnahme des Vertragsarztes durch einen Patienten 1.7 Gesundheits- und Früherkennungsuntersuchungen, Mutterschaftsvorsorge, Empfängnisregelung und Schwangerschaftsabbruch 2.1 Infusionen, Transfusionen, Retransfusionen, Programmierung von Medikamentenpumpen 2.3 Kleinchirurgische Eingriffe, Allgemeine therapeutische Leistungen
III	Arztgruppenspezifische Gebührenordnungspositionen	3.2 Gebührenordnungspositionen der allgemeinen hausärztlichen Versorgung 4.2 Gebührenordnungspositionen der Kinder- und Jugendmedizin 6.2 Augenärztliche Grundpauschalen 7.2 Chirurgische Grundpauschalen 23.2 Psychotherapeutische Grundpauschalen
IV	Arztgruppenübergreifende spezielle Gebührenordnungspositionen	31.2 ambulante Operationen 31.3 postoperative Überwachungskomplexe 33 Ultraschalldiagnostik 34.2 Diagnostische Radiologie
V	Kostenpauschalen	40.3 Kostenpauschalen für Versandmaterial, Versandgefäße usw. sowie für die Versendung/Transport von Untersuchungsmaterial, Röntgenaufnahmen und Filmfolien 40.4 Kostenpauschalen für die Versendung/Transport von Briefen, schriftlichen Unterlagen, Kostenpauschalen für Telefax
VI	Anhänge	
VII	Ausschließlich im Rahmen der ambulanten spezialfachärztlichen Versorgung (ASV) berechnungsfähige Gebührenordnungspositionen	
VIII	Ausschließlich im Rahmen von Erprobungsverfahren gemäß §137e SGB V berechnungsfähige Gebührenordnungspositionen	

1.7.4 PZN

Die **Pharmazentralnummer (PZN)** ist eine achtstellige Nummer, die jedes Arzneimittelprodukt, also jedes Handelspräparat, das in Deutschland auf dem Markt ist, unter Differenzierung der Zubereitungsart und Packungsgröße kennzeichnet. Auch Medizinprodukte und andere apothekenübliche Produkte haben i.d.R. eine PZN. Die PZN ist notwendig für die Abrechnung mit den Krankenkassen und wird daher auch immer zu Abrechnungszwecken von den Apotheken an die Krankenkassen gemeldet. Da

die PZN als fortlaufende Nummer an neue Produkte vergeben wird, liefert sie selbst keine inhaltliche Information. Man kann aus der PZN also z. B. nicht direkt die Wirkgruppe oder den Wirkstoff ableiten. Insofern bildet die PZN-Systematik auch kein Klassifikationssystem. Erst durch die Verknüpfung der PZN mit dem durch sie bezeichneten Produkt gelangt man an die Informationen zu Wirkstoffen, Darreichungsform, Hersteller, Menge/Art der Dosierung und Preis. In den vergangenen Jahren wurden PZN von Produkten, die nicht mehr auf dem Markt sind, nach einer Wartezeit von einigen Jahren neuvergeben. Bei Analysen über mehrere Jahre hinweg muss also beachtet werden, dass eine PZN ggf. unterschiedliche Produkte kennzeichnen kann.

1.7.5 ATC und DDD

Anatomisch-Therapeutisch-Chemisches System (ATC)

Mithilfe des **Anatomisch-Therapeutisch-Chemischen Klassifikationssystems (ATC)** werden Arzneimittel klassifiziert. Die Klassifikation gilt für die Substanzen bzw. Wirkstoffe, nicht für die Handelspräparate. Für den deutschen Arzneimittelmarkt gibt das Bundesinstitut für Arzneimittel und Medizinprodukte jährlich die aktuelle ATC-Klassifikation heraus, die eine Anpassung an die von der WHO herausgegebene ATC-Klassifikation ist und vom Wissenschaftlichen Institut der AOK (WidO) erstellt wird. Die aktuelle Version der WHO-ATC-Klassifikation kann auf den Seiten des WHO Collaborating Centre for Drug Statistics Methodology eingesehen werden (unter: whocc.no/atc_ddd_index/, Stand 12/2024), die deutsche Version auf den Seiten des Bundesinstituts für Arzneimittel und Medizinprodukte (https://www.bfarm.de/DE/Kodiersysteme/Klassifikationen/ATC/_node.html, Stand 12/2024). Im deutschen ATC-System sind u.a. auch pflanzliche und homöopathische Zubereitungen ergänzt.

Das ATC-System ist hierarchisch aufgebaut und besteht aus fünf Ebenen, die die Arzneimittel zunehmend stärker unterteilen und differenzieren. Die oberste Ebene ist in 15 anatomische Hauptgruppen unterteilt (s. Tabelle 1-18), die das Organ oder System bezeichnen, auf die das Arzneimittel wirkt und die jeweils mit einem Buchstaben gekennzeichnet sind (z. B. Gruppe **B:** Blut und blutbildende Organe). Auf der nächsten Ebene wird die therapeutisch-pharmakologische Untergruppe spezifiziert. Dazu werden zwei Ziffern an den Buchstaben der Hauptgruppe gehängt (z. B. B**01:** Antithrombotische Mittel). Auf dieser dreistelligen Ebene der ATC-Codes lassen sich somit mehr als 100 verschiedene Kategorien unterscheiden. Die dritte und vierte Ebene ist nach chemischen, therapeutischen oder pharmakologischen Eigenschaften geordnet. Zur Differenzierung wird der ATC-Code um jeweils einen Buchstaben pro Ebene verlängert (z. B. B01**AC:** Thrombozytenaggregationshemmer, excl. Heparin). Auf der fünften Ebene gibt es Untergruppen für die chemische Substanz. Der ATC-Code wird an dieser Stelle um zwei Ziffern für die chemische Substanz erweitert (z. B. B01AC**06:** Acetylsalicylsäure). Durch die zunehmende Differenzierung auf den unteren Ebenen lässt sich eine große Anzahl an Substanzen systematisch ordnen. Auf der feinsten Gliederungsebene (also im Bereich der 7-stelligen Codes) existieren mehr als 6.500 verschiedene Codes.

Ein Wirkstoff kann je nach Anwendung und Dosierung unterschiedliche ATC-Codes erhalten. Acetylsalicylsäure hat z. B. als Thrombozytenaggregationshemmer einen anderen Code als Schmerzmittel.

Tabelle 1-18. Kapitel des Anatomisch-Therapeutisch-Chemischen Klassifikationssystems (ATC)

Kapitel	Gliederung	Titel
ATC A	A01-A16	Alimentäres System und Stoffwechsel
ATC B	B01-B06	Blut und blutbildende Organe
ATC C	C01-C10	Kardiovaskuläres System
ATC D	D01-D11	Dermatika
ATC G	G01-G04	Urogenitalsystem und Sexualhormone

ATC H	H01-H05	Hormone, systemisch (ohne Sexualhormone)
ATC J	J01-J07	Antiinfektiva für systemische Gabe
ATC L	L01-L04	Antineoplastische und immunmodulierende Substanzen
ATC M	M01-M09	Muskel- und Skelettsystem
ATC N	N01-N07	Nervensystem
ATC P	P01-P03	Antiparasitäre Substanzen, Insektizide, Repellenzien
ATC Q	…	Veterinärmedizinische Arzneimittel
ATC R	R01-R07	Respirationstrakt
ATC S	S01-S03	Sinnesorgane
ATC V	V01-V90	Verschiedene

Defined daily dose (DDD)

Mit der ATC-Klassifikation eng zusammen hängt die **definierte Tagesdosis (defined daily dose, DDD)** eines Arzneimittels. Eine DDD beschreibt die typischerweise täglich erforderliche Verabreichungsdosis eines Wirkstoffs in der Hauptindikation für einen durchschnittlichen Erwachsenen. Wenn Arzneimittel hauptsächlich Kindern verordnet werden, dann gibt es ggf. auch Angaben zu Kinderdosierungen. DDDs werden wirkstoffabhängig festgelegt und i.d.R. auch mit der ATC-Klassifikation veröffentlicht (dort dann zu den 7-stelligen Codes, mit denen i.d.R. einzelne Wirkstoffe spezifiziert sind). Beispielsweise wird für den Betablocker Propranolol mit dem ATC-Code C07AA05 als DDD (bei oraler Einnahme) die Menge von 0,16 Gramm angegeben, die dann für die Behandlung eines Bluthochdrucks mit dem Betablockeran einem Tag ausreichen sollte.

Bezug zwischen PZN und ATC-Code

(Fertig-)Arzneimittel lassen sich in Deutschland, wie bereits erläutert, eindeutig über die PZN identifizieren. Zu einem überwiegenden Teil dieser PZN existieren auch ATC- und DDD-Zuordnungen aus unterschiedlichen Quellen, wobei im GKV-Bereich zumeist Zuordnungen verwendet werden, die (wie die ATC-Klassifikation selbst) vom Wissenschaftlichen Institut der AOK (WidO) gepflegt und monatlich aktualisiert werden. Ist mit einer PZN beispielsweise die Verordnung eines Präparats gekennzeichnet, das 100 Filmtabletten mit je 80 mg Wirkstoff Propranolol enthält, sollte dann diesem Präparat der ATC-Code C07AA05 und die Mengenangabe 50 DDD zugeordnet sein, da die Packung 100 x 0,08 g Propranolol enthält und 0,16 g pro Tag bei typischer Behandlung erforderlich sind. Über die Zuordnung des ATC-Codes mit DDD-Angaben zur PZN in den betrachteten Daten lässt so ableiten, dass ein Betablocker verordnet wurde, der in gewöhnlicher Dosierung für eine Behandlung über 50 Tage ausreichen sollte.

1.8 Kennzeichnungen von Personen und Einrichtungen im Gesundheitssystem

Relevante Beobachtungseinheiten werden in den Krankenkassendaten mit Identifikationsnummern gekennzeichnet, die für die Krankenkassen die genaue Identifikation der Beobachtungseinheit, also z. B. einer Person oder einer Betriebsstätte, ermöglichen. Aus Datenschutzgründen werden diese Nummern für Forschungszwecke aber i.d.R. pseudonymisiert.

Krankenversichertennummer
Zur Identifikation der Versicherten ist die Krankenversichertennummer notwendig (§290 SGB V). Über sie können alle Versicherten eindeutig identifiziert werden und in Anspruch genommene Leistungen können den Versicherten zugeordnet werden. Sie besteht aus einem unveränderbaren Teil, über den die Versicherten auch krankenkassenübergreifend identifiziert werden können, sowie einem veränderbaren Teil, der kassenspezifische Informationen erhält.

Lebenslange Arztnummer (LANR)
Die Lebenslange Arztnummer (LANR) ist eine neunstellige Nummer, die die eindeutige Identifikation eines Arztes/einer Ärztin oder eines Psychotherapeuten/einer Psychotherapeutin ermöglicht. Die Nummer wird von der Kassenärztlichen Vereinigung an alle Ärzte/Ärztinnen und Psychotherapeuten/Psychotherapeutinnen vergeben, die an der vertragsärztlichen Versorgung teilnehmen, also ambulante Leistungen mit der gesetzlichen Krankenversicherung abrechnen dürfen. Die Nummer besteht ein Leben lang, sie ist also auch unabhängig vom Tätigkeitsort oder vom Status. Wenn eine Person mehreren Fachgruppen angehört, können der Person auch mehrere LANR zugeordnet werden. Da die letzten beiden Ziffern der LANR die Fachgruppe kodieren, sind in einem solchen Fall lediglich die letzten zwei Ziffern unterschiedlich, die ersten sieben Ziffern bleiben gleich.

Betriebsstättennummer (BSNR)
Die Betriebsstättennummer (BSNR) ist eine neunstellige Nummer, die eindeutig den Ort (also die Praxis oder das medizinische Versorgungszentrum) der Leistungserbringung kennzeichnet. Eine BSNR erhalten alle vertragsärztlich abrechnenden Praxen. Krankenhäuser bekommen nur dann eine BSNR, wenn sie bestimmte Leistungen erbringen. Die BSNR ist an den Praxisstandort gebunden, nicht an die dort tätigen Personen. Jeder Abrechnungsfall aus der ambulanten Versorgung ist genau einer BSNR und damit genau einer Praxis zugeordnet, wobei innerhalb dieser Praxis dann ggf. auch mehrere Ärzte/Ärztinnen an der abgerechneten Behandlung beteiligt sein können.

Institutionskennzeichen (IK)
Das Institutionskennzeichen (IK) ist eine neunstellige Nummer, die eindeutig Institutionen und Leistungserbringer im Gesundheitssystem kennzeichnet. Alle Leistungserbringer, die im Gesundheitswesen mit den Krankenkassen und anderen Sozialversicherungen Leistungen abrechnen, können ein IK beantragen, also z. B. Krankenhäuser, Arzt- und Zahnarztpraxen, medizinische Versorgungszentren, Labore, Apotheken, Hebammen oder Krankentransportunternehmen. Beim neunstelligen IK geben die ersten zwei Ziffern Auskunft darüber, um welche Art von Leistungserbringer es sich handelt. Die 3. Und 4. Ziffer gibt das Bundesland an. In Routinedaten bei Krankenkassen werden insbesondere Krankenhäuser und Rehabilitationseinrichtungen regelmäßig eindeutig durch das IK gekennzeichnet, wohingegen bei niedergelassenen Ärzten/Ärztinnen eine Zuordnung über BSNR und LANR erfolgt.

1.9 Besonderheiten von Krankenkassenroutinedaten in der Forschung

Bei der Routinedatenanalysen muss bedacht werden, dass ihre Entstehung im Rahmen von Abrechnungsprozessen zu ganz eigenen Besonderheiten führt, die ihre Verwendung für die Forschung beeinflussen (Schubert et al., 2008).

Validität

Da Routinedaten im Rahmen der Abrechnungsprozesse entstehen, muss die Validität von Routinedaten je nach Forschungsfrage unterschiedlich bewertet werden. So wird das Versorgungsgeschehen im GKV-Bereich naturgemäß recht valide abgebildet, abgesehen beispielsweise von möglichen Datenfehlern oder falschen Angaben. Bei anderen Forschungsfragen, z. B. zu Erkrankungshäufigkeiten kann jedoch nur von einer eingeschränkten Validität der Routinedaten ausgegangen werden. Dies ergibt sich v.a., weil die Daten ereignisbezogen erfasst werden, d. h. sie entstehen nur, wenn die Versicherten das Gesundheitssystem in Anspruch nehmen und deshalb eine Leistung bei den Krankenkassen geltend gemacht wird. Wenn aber die Versicherten z. B. eine Erkrankung ohne Arzt-/Krankenhausbesuch auskurieren (wie z. B. häufig bei einer Erkältung oder leichten Grippe), dann tauchen auch keine Informationen dazu in den Daten der Krankenkassen auf.

Aber auch wenn Versicherte das Gesundheitssystem in Anspruch nehmen, dann entstehen Daten nur auf Seiten der Leistungserbringer, es gibt aber i.d.R. keine Information darüber, wie die Versicherten sich verhalten. So kann es sein, dass Medikamente, die in den Apotheken ausgegeben werden (und damit in den Abrechnungsdaten auftauchen) von den Versicherten gar nicht eingenommen werden (Horenkamp-Sonntag, Lindner, Wenzel, Gerste, & Ihle, 2014). Auch über das weitere Gesundheitsverhalten der Versicherten (z. B. Ernährung oder Sport) ist i.d.R. nichts bekannt, was aber dennoch einen Einfluss auf die weitere Inanspruchnahme des Gesundheitswesens hat.

Zufällige und systematische Fehler

Wie bei allen Daten, die erhoben werden, ist es auch bei den Routinedaten möglich, dass sie fehlerbehaftet sind, z. B. aufgrund einer fehlerhaften Eingabe oder Übertragung. Bei den Arzneimitteldaten wird z. B. häufig ein Papierrezept eingescannt, was u.U. zu Schwierigkeiten bei der Zeichenerkennung führt (Horenkamp-Sonntag et al., 2014).

Abgesehen von solchen, mehr oder weniger zufälligen, Fehlern kann aber auch eine systematische Fehlerhaftigkeit in den Daten stecken. Da die Daten für die Leistungsabrechnung relevant sind, könnten abrechnungsrelevante Diagnosen überdokumentiert werden oder die Schwere von Krankheiten überschätzt werden. In bestimmten Bereichen kann es aber auch zu einer Unterschätzung von Diagnosen kommen. So werden psychische und Verhaltensstörungen (Kapitel F im ICD) ggf. als somatische Diagnose (z. B. Rückenschmerz) gespeichert, um eine Stigmatisierung zu vermeiden (Horenkamp-Sonntag et al., 2014).

Vollständigkeit

Die Datenvollständigkeit versteht sich als Ausmaß von vorhandenen bzw. fehlenden Datenelementen oder Attributen in einem Datensatz und spielt in der empirischen Forschung eine wichtige Rolle für die Korrektheit und Repräsentativität von Forschungsergebnissen. Problematisch ist hierbei vor allem das gehäufte Auftreten von fehlenden Werten (missings). Fehlende Werte führen zu einer Reduktion der Fallzahl, die für eine bestimmt Analyse zur Verfügung steht. Viele statistische Analysen setzen dabei vollständige Datensätze voraus und berücksichtigen nur Fälle mit kompletten Beobachtungen ohne fehlende Werte. In der Folge gilt, dass je höher der Merkmalsumfang und der Anteil an fehlenden Werten ausfallen, desto größer ist die erwartbare Zahl auszuschließender Fälle. Darüber hinaus kön-

nen Parameterschätzungen verzerrt sein, wenn systematische Unterschiede zwischen den beobachteten und den fehlenden Daten bestehen und fehlende Werte nicht rein zufällig in den Daten verteilt sind (engl. missing completely at random, MCAR).

An dieser Stelle stellt sich für Forschende die Frage, welche Datenvollständigkeit in den Routinedaten bei Krankenkassen erwartbar ist und ob es Merkmale gibt, die häufiger von fehlenden Werten betroffen sind. Grundsätzlich werden die Abrechnungsdaten von den Krankenkassen im Rahmen ihrer gesetzlichen Aufgaben umfangreich geprüft, damit sie für den primären Zweck der Abrechnung verwendet werden können. Dieses Prozedere schließt eine Prüfung unvollständig dokumentierter Daten mit ein, so dass alle abrechnungsrelevanten Daten weitgehend vollständig vorliegen. Dies gilt auch für Routinedaten, die für Forschungszwecke bereitgestellt werden, da in der Regel nur abgeschlossene und geprüfte Fälle von der Krankenkasse an die Forschungseinrichtungen übermittelt werden. Demnach ist festzuhalten, dass fehlende Werte kaum ein Problem bei der Analyse von Routinedaten im Forschungskontext darstellen. Da Routinedaten ereignisbezogen erfasst werden, sind fehlende Dokumentationen bei nachweislich vorhandenem Beobachtungszeitraum zumeist nicht auf das Vorliegen fehlender Werte im engeren Sinn zurückzuführen, sondern darauf, dass kein Ereignis stattgefunden hat. Dementsprechend sind in der Regel keine komplexen Methoden zur Identifikation (z. B. mittels Musteranalyse, MCAR-Test nach Little) und Imputation von fehlenden Werten (z. B. mittels Mittelwerts- oder multipler Imputation) erforderlich.

Gleichwohl gibt es in bestimmten Fällen Merkmale, die von den Krankenkassen für einen konkreten Beobachtungszeitraum nicht bereitgestellt werden können. Dies ist oftmals dadurch bedingt, dass bestimmte Daten von den Krankenkassen lediglich für den Zeitraum der Abrechnung in aktueller Form vorgehalten werden müssen. Da beispielsweise nur die aktuelle Anschrift des Versicherten benötigt wird, besteht derzeit im Falle eines Wohnortwechsels eine sehr heterogene Praxis in der Archivierung der alten Anschrift bzw. von Elementen der Anschrift, wie der Postleitzahl. Das Fehlen einer Postleitzahl-Historie wäre zum Beispiel bei der Planung von Forschungsvorhaben zu berücksichtigen, die eine raum- und zeitbezogene Auswertung von historischen Routinedaten anhand der Wohnort-Postleitzahl vorsehen (Beispiel: Abschätzung von Hautkrebsinzidenzen in den letzten 10 Jahren auf Ebene 5-stelliger Postleitzahlgebiete). Darüber hinaus gibt es auch Merkmale, die zu einem aktuellen Stichtag oder Zeitraum zwar nicht verfügbar sind, sich unter Umständen jedoch aus Informationen aus anderweitigen Zeiträumen rekonstruieren lassen. Ein Beispiel hierfür wäre die Angabe zum Ausbildungsabschluss, die bei Rentnerinnen und Rentnern fehlt, sich aber aus Daten aus dem Zeitraum vor Renteneintritt herleiten lässt. Werden entsprechende Daten für ein Forschungsvorhaben benötigt, empfiehlt sich eine frühzeitige Kontaktaufnahme mit der projektteilnehmenden Krankenkasse, um die Datenverfügbarkeit in Erfahrung zu bringen und Möglichkeiten einer Bereitstellung auszuloten.

Repräsentativität

Die Daten einzelner Krankenkassen können zumeist nicht als repräsentativ für die Gesamtpopulation (z. B. deutschlandweit) betrachtet werden (Grobe & Ihle, 2014), d. h. die Versichertenpopulationen unterscheiden sich zwischen den Krankenkassen und im Vergleich zur Allgemeinbevölkerung, zum Teil aus historischen Gründen. Auch bleibt die Versichertenpopulation einer Krankenkasse nicht gleich, sondern bildet eine dynamische Kohorte, da Ein- und Austritte im Prinzip jederzeit möglich sind. Unter Umständen können die Routinedaten jedoch in ihrer Repräsentativität sogar Primärdaten übertreffen, z. B. wenn es um Zustände oder Krankheiten geht, die in einer Primärerhebung eher verschwiegen werden würden, wie z. B. ein Krankenhausaufenthalt aufgrund von Alkoholintoxikation (Grobe & Ihle, 2014).

Aussagen über Erkrankungshäufigkeiten

Da in den Routinedaten auch die ermittelten Diagnosen enthalten sind, lassen sich prinzipiell auch Aussagen über Erkrankungshäufigkeiten treffen (z. B. Köster et al., 2004; Petersen, Wittmann, Arndt, & Göpffarth, 2014). Allerdings kann man die Diagnoseraten i.d.R. nicht direkt mit den Erkrankungshäu-

figkeiten gleichsetzen. Dafür gibt es mehrere Gründe: Zum einen werden Diagnoseraten dadurch beeinflusst, ob Betroffene das Gesundheitssystem überhaupt in Anspruch nehmen (können). Nur dann können entsprechende Diagnosen codiert werden. Zum anderen kann es im Behandlungsalltag z. B. zu fehlerhaften Kodierungen von Diagnosen kommen, insbesondere wenn die Dokumentation wenig standardisiert ist. Trotzdem können die Diagnoseraten zumindest als Anhaltspunkt dienen oder einen Vergleich unterschiedlicher Subgruppen ermöglichen (bei denen dann von einem ähnlichen Anteil an Fehlklassifizierungen ausgegangen werden kann (Grobe & Dräther, 2014; Schubert, Ihle, & Köster, 2010)).

Datenverzug

Bei allen Daten der Krankenkassen kommt es zu einem gewissen zeitlichen Verzug zwischen der Entstehung der Daten (also dem Arzt-/Krankenhausbesuch, der Arzneimittelausgabe etc.) und dem Zeitpunkt, zu dem die Daten bei den Krankenkassen vollständig vorliegen. Je nach Bereich des Gesundheitssystems ergeben sich unterschiedliche Dauern, es ist aber immer mit mindestens ein paar Monaten zu rechnen. Durch die quartalsweise Übermittlung der Daten im ambulanten Bereich zunächst an die KVen und von dort zu den Krankenkassen kommt es i.d.R. zu einem längeren Datenverzug, bis die Daten bei den Krankenkassen abgerufen werden können. So liegen die vollständigen ambulanten Daten für ein Jahr teilweise erst zu Beginn der zweiten Hälfte des Folgejahres vor (Grobe & Dräther, 2014). Der Zeitpunkt, zu dem die Daten aus Krankenhausbehandlungen vollständig und korrekt bei der Krankenkasse vorliegen, lässt sich nicht einfach voraussagen. Zum einen dauert es einige Zeit nach der Entlassung der Patienten bis die stationären Daten vollständig vorliegen. So empfiehlt sich eine Wartezeit von mindestens drei Monaten nach Entlassdatum (Grobe et al., 2014). Zum anderen unterscheidet sich aber auch die Liegedauer im Krankenhaus von Fall zu Fall teilweise deutlich. d. h. ein Teil der Patienten verbleibt so lange im Krankenhaus, dass die Daten bei der Krankenkasse sehr lange unvollständig bleiben.

Behäbigkeit von Klassifikationssystemen und des GKV-Abrechnungswesens

Die Standardisierung der Datenübermittlung im Gesundheitssystem erfordert, dass für alle Situationen Regeln bestehen, wie diese Situationen erfasst werden müssen. Dies resultiert in einer gewissen Behäbigkeit, was neue Entwicklungen und komplexe Situationen angeht und führt dazu, dass gerade neue Entwicklungen und Erkrankungen, sowie komplexe Syndrome oft nicht adäquat erfasst werden können. Als aktuelles Beispiel dient z. B. die Covid-19-Erkrankung oder auch Long Covid, die nach ihrem Auftreten nachträglich im aktuellen ICD-10-GM eingepflegt werden mussten. Auch komplexe Syndrome wie bei einem Burnout lassen sich nicht unmittelbar über einen ICD-Schlüssel abbilden.

Trotz der genannten Besonderheiten und Einschränkungen bieten Routinedaten eine wertvolle Basis für gesundheitsbezogene Fragestellungen. Unter Berücksichtigung der Eigenheiten dieser Daten können Routinedaten somit für eine große Bandbreite an Forschungsvorhaben genutzt werden.

Weiterführende Literatur

Thema	Quellen
Tiefergehende Einführung in Routinedaten des Gesundheitswesens	– Swart, E., Ihle, P., Gothe, H., & Matusiewicz, D. (Eds.). (2014). Routinedaten im Gesundheitswesen. Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven (Vol. 2., vollst. überarbeitete Ausgabe). Bern: Hans Huber. – Neubauer, S., Zeidler, J., Lange, A., & Graf von der Schulenburg, J.-M. (2017). Prozessorientierter Leitfaden für die Analyse und Nutzung von Routinedaten der Gesetzlichen Krankenversicherung (1 ed.). Baden-Baden: Nomos Verlagsgesellschaft mbH & Co. KG. – Ohlmeier, C; Frick, J; Prütz, F; Lampert, T; Ziese, T; Mikolajczyk, R; Garbe, E (2014). Nutzungsmöglichkeiten von Routinedaten der Gesetzlichen Krankenversicherung in der Gesundheitsberichterstattung des Bundes. Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz 57(4): 464-472.
Detaillierte Ausführung, welche Daten von den Leistungserbringern an die Krankenkassen übermittelt werden und Schlüsselverzeichnisse für die übermittelten Daten	– Technische Anlagen (TA) der Richtlinien des Datenaustauschverfahrens zwischen Leistungserbringern und gesetzlicher Krankenversicherung: https://www.gkv-datenaustausch.de/
Beispielprojekte für die Nutzung von Routinedaten	– Krüger-Brand, HE (2014). Projekt der Knappschaft: Elektronische Behandlungsinformation. Deutsches Ärzteblatt 111(31-32): 1379. – Weinand, S; Thürmann, PA; Dröge, P; Koetsenruijter, J; Klora, M; Grobe, TG (2022). Potentiell inadäquate Medikation bei Heimbewohnern: Eine Analyse von Risikofaktoren anhand bundesweiter GKV-Routinedaten der AOK für das Jahr 2017. Gesundheitswesen 84(5): 448-456. – Grobe, TG; Kleine-Budde, K; Bramesfeld, A; Thom, J; Bretschneider, J; Hapke, U (2019). Prävalenzen von Depressionen bei Erwachsenen – eine vergleichende Analyse bundesweiter Survey- und Routinedaten. Gesundheitswesen 81(12): 1011-1017.

Quellen

Bundesministerium für Gesundheit. (2021). Gesetzliche Krankenversicherung. Mitglieder, mitversicherte Angehörige und Krankenstand. Jahresdurchschnitt 2021. Retrieved from https://www.bundesgesundheitsministerium.de/themen/krankenversicherung/zahlen-und-fakten-zur-krankenversicherung/mitglieder-und-versicherte.html

BARMER Institut für Gesundheitssystemforschung (bifg) (2024). Hilfsmittelreport 2023. Retrieved from https://www.bifg.de/media/dl/Reporte/Heil-und-Hilfsmittelreporte/2023/barmer-hilfsmittelreport-2023-neu.pdf

Gemeinsamer Bundesausschuss. (2022). *Richtlinie des Gemeinsamen Bundesausschusses über die Verordnung von Arzneimitteln in der vertragsärztlichen Versorgung*. BAnz AT 05.07.2022 B1 Retrieved from https://www.g-ba.de/richtlinien/3/.

Gesundheitsberichtserstattung des Bundes. (2022a). Definition: Krankenhäuser. Retrieved from https://www.gbe-bund.de/gbe/pkg_isgbe5.prc_show_dokument?p_aid=48601791&p_uid=gast&sprache=D&p_lfd_nr=1&p_dokumente=1

Gesundheitsberichtserstattung des Bundes. (2022b). Haupt- und Nebendiagnose. Retrieved from https://www.gbe-bund.de/gbe/pkg_isgbe5.prc_show_dokument?p_aid=55831317&p_uid=gast&sprache=D&p_lfd_nr=1&p_dokumente=1

Grobe, T. G., & Dräther, H. (2014). Ambulante ärztliche Versorgung. In E. Swart, P. Ihle, H. Gothe, & D. Matusiewicz (Eds.), *Routinedaten im Gesundheitswesen. Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven* (Vol. 2). Bern: Hans Huber.

Grobe, T. G., & Ihle, P. (2014). Stammdaten und Versichertenhistorien. In E. Swart, P. Ihle, H. Gothe, & D. Matusiewicz (Eds.), *Routinedaten im Gesundheitswesen. Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven* (Vol. 2). Bern: Hans Huber.

Grobe, T. G., Nimptsch, U., & Friedrich, J. (2014). Krankenhausbehandlung. In E. Swart, P. Ihle, H. Gothe, & D. Matusiewicz (Eds.), *Routinedaten im Gesundheitswesen. Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven* (Vol. 2). Bern: Hans Huber.

Horenkamp-Sonntag, D., Lindner, R., Wenzel, F., Gerste, B., & Ihle, P. (2014). Prüfung der Datenqualität und Validität von GKV-Routinedaten. In E. Swart, P. Ihle, H. Gothe, & D. Matusiewicz (Eds.), *Routinedaten im Gesundheitswesen. Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven* (Vol. 2). Bern: Hans Huber.

Institut für das Entgeltsystem im Krankenhaus. (2022). Deutsche Kodierrichtlinien 2022. Retrieved from https://www.g-drg.de/aG-DRG-System_2022/Kodierrichtlinien/Deutsche_Kodierrichtlinien_2022

Klemm, A.-K., Knieps, F., Pfaff, H. (Hrsg.) (2024). BKK Gesundheitsreport 2024 - Spurwechsel Prävention. Berlin: MWV Medizinisch Wissenschaftliche Verlagsgesellschaft.

Köster, I., Schubert, I., Döpfner, M., Adam, C., Ihle, P., & Lehmkuhl, G. (2004). Hyperkinetische Störungen bei Kindern und Jugendlichen: Zur Häufigkeit des Behandlungsanlasses in der ambulanten Versorgung nach den Daten der Versichertenstichprobe AOK Hessen/KV Hessen (1998-2001). *Zeitschrift für Kinder-und Jugendpsychiatrie und Psychotherapie, 32*(3), 157-166.

Motzek, T., Werblow, A., Schmitt, J., & Marquardt, G. (2019). Administrative Prävalenz und Versorgungssituation der Demenz im Krankenhaus–Eine versorgungsepidemiologische Studie basierend auf GKV-Daten sächsischer Versicherter. *Das Gesundheitswesen, 81*(12), 1022-1028.

Neubauer, S., Zeidler, J., Lange, A., & Graf von der Schulenburg, J.-M. (2017). *Prozessorientierter Leitfaden für die Analyse und Nutzung von Routinedaten der Gesetzlichen Krankenversicherung* (1 ed.). Baden-Baden: Nomos Verlagsgesellschaft mbH & Co. KG.

Petersen, G., Wittmann, R., Arndt, V., & Göpffarth, D. (2014). Epidemiologie der Multiplen Sklerose in Deutschland. *Der Nervenarzt, 85*(8), 990-998.

Schmitt, N., & Wende, D. (2021). *Heilmittelreport 2021*. Retrieved from https://www.barmer.de/presse/infothek/studien-und-reporte/heil-und-hilfsmittelreport

Schmitt, N., & Wende, D. (2022). *Hilfsmittelreport 2022*. Retrieved from https://www.barmer.de/presse/infothek/studien-und-reporte/heil-und-hilfsmittelreport

Schröder, H. (2014). Arzneimittelverordnungen. In E. Swart, P. Ihle, H. Gothe, & D. Matusiewicz (Eds.), *Routinedaten im Gesundheitswesen. Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven* (Vol. 2). Bern: Hans Huber.

Schubert, I., Ihle, P., & Köster, I. (2010). Interne Validierung von Diagnosen in GKV-Routinedaten: Konzeption mit Beispielen und Falldefinition. *Das Gesundheitswesen, 72*(06), 316-322.

Schubert, I., Köster, I., Küpper-Nybelen, J., & Ihle, P. (2008). Versorgungsforschung mit GKV-Routinedaten. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz, 51*(10), 1095-1105. doi:10.1007/s00103-008-0644-0

Schwinger, A., Behrendt, S., Tsiasioti, C., Stieglitz, K., Breitkreuz, T., Grobe, T. G., & Klauber, J. (2018). Qualitätsmessung mit Routinedaten in deutschen Pflegeheimen: Eine erste Standortbestimmung. In *Pflege-Report 2018* (pp. 97-125): Springer, Berlin, Heidelberg.

Spitzenverband Bund der Krankenkassen (GKV-Spitzenverband) (2024). GKV-Heilmittel-Schnellinformation für Deutschland nach § 84 Abs. 5 i.V.m. Abs. 7 SGB V. Januar bis Dezember 2023. Retrieved from https://www.gkv-heilmittel.de/media/dokumente/his_statistiken/2023_04/Bundesbericht_HIS-Bericht_202304.pdf

Swart, E., Deh, U., & Robra, B.-P. (2008). Die Nutzung der GKV-Daten für die kleinräumige Analyse und Steuerung der stationären Versorgung. *Bundesgesundheitsblatt-Gesundheitsforschung-Gesundheitsschutz, 51*(10), 1183-1192.

Swart, E., & Ihle, P. (2008). Der Nutzen von GKV-Routinedaten für die Versorgungsforschung. In (Vol. 51, pp. 1093-1094): Springer.

Swart, E., Ihle, P., Gothe, H., & Matusiewicz, D. (Eds.). (2014). *Routinedaten im Gesundheitswesen. Handbuch Sekundärdatenanalyse: Grundlagen, Methoden und Perspektiven* (Vol. 2., vollst. überarbeitete Ausgabe). Bern: Hans Huber.

2 Abgrenzung und Darstellung relevanter KI-Analysetechniken

Im Gesundheitswesen finden sich verschiedenste Anwendungen von KI, sei es bei der Analyse von komplexen medizinischen Daten oder der Erkennung von Tumoren bei bildgebenden Verfahren. Meistens handelt es sich in der Medizin um sensible Daten, welche durch Datenschutzmaßnahmen einen besonderen Umgang benötigen. Im folgenden Kapitel werden die verschiedenen Arten Maschinellen Lernens, einem Teilbereich der KI, erklärt. Insbesondere wird auf einige Methoden des überwachten Lernens (Supervised Learning) genauer eingegangen, da diese in Kapitel 5 beispielhaft an Routinedaten angewendet werden Hierzu gehören mitunter die logistische Regression, Random Forests, Adaptive Boosting und künstliche Neuronale Netze. Darauffolgend werden einige Prinzipien der Validierung und Optimierung dieser Methoden erläutert. Zudem werden Möglichkeiten beschrieben, wie mit Daten umgegangen werden kann, die ein Ungleichgewicht verschiedener Klassen aufweisen. Zuletzt wird auf die Interpretierbarkeit und insbesondere erklärbare KI (XAI) eingegangen, da dies bei einigen KI-Methoden durchaus eine Herausforderung darstellen kann.

Der Weg der Künstlichen Intelligenz (KI) im Gesundheitswesen begann mit der Entwicklung von regelbasierten Expertensystemen in den 1970er und 1980er Jahren, die die Entscheidungsfähigkeit menschlicher Experten nachahmen sollten (Leondes, 2002). Seitdem hat sich der Bereich erheblich weiterentwickelt, vor allem durch Fortschritte bei der Rechenleistung, die Verfügbarkeit großer Datensätze und Verbesserungen bei den Algorithmen. Das maschinelle Lernen (ML), ein Teilbereich der KI, der sich auf die Entwicklung von Modellen konzentriert, die von Daten lernen, hat sich in vielen Anwendungsbereichen zunehmend etabliert. ML-Modelle werden vermehrt zur Analyse komplexer medizinischer Daten eingesetzt, zum Beispiel in der Analyse verzerrter Ergebnisdaten zur gesundheitsbezogenen Lebensqualität (HRQOL) (Norris et al., 2006) oder zur Vorhersage des bildgebend bestätigten Wiederauftretens von Prostatakrebs (J. M. Beinecke et al., 2022). Vor allem Deep-Learning (DL)-Modelle, eine Unterklasse von ML-Modellen, finden auf Grund ihrer hohen Vorhersagegenauigkeiten, wie sie insbesondere für klinische Anwendungen erforderlich sind, in den letzten Jahren zunehmend Anwendung (Jumper et al., 2021; Shigemizu et al., 2023; Viros et al., 2008). ML-Modelle können angepasste Behandlungsprotokolle vorschlagen, die die Erfolgswahrscheinlichkeit auf der Grundlage individueller genetischer Marker, Medikation, Lebensstilfaktoren sowie früherer Behandlung optimieren. Durch den

Einsatz von Algorithmen zur Erkennung von Mustern und deren Fähigkeit, aus einer großen Menge von Gesundheitsdaten zu lernen, können KI-Modelle das Sterblichkeitsrisiko (Schünemann et al., 2000) und die Wahrscheinlichkeit der Entstehung oder Weiterentwicklung von Krankheiten bei Patienten vorhersagen (Jamshidi et al., 2019). Die Vorhersagefähigkeiten von ML-Modellen können Krankenhäuser beim effektiven Ressourcenmanagement unterstützen (Karaarslan & Aydın, 2021). KI-gesteuerte Roboter werden zum Beispiel in der Chirurgie eingesetzt, um die Präzision zu erhöhen, die Operationszeiten zu verkürzen und die Genesungszeiten zu minimieren. Diese Systeme nutzen Echtzeitdaten, um Chirurgen zu assistieren und können bestimmte Handlungen autonom und mit hoher Präzision durchführen (Mathis-Ullrich & Scheikl, 2021).

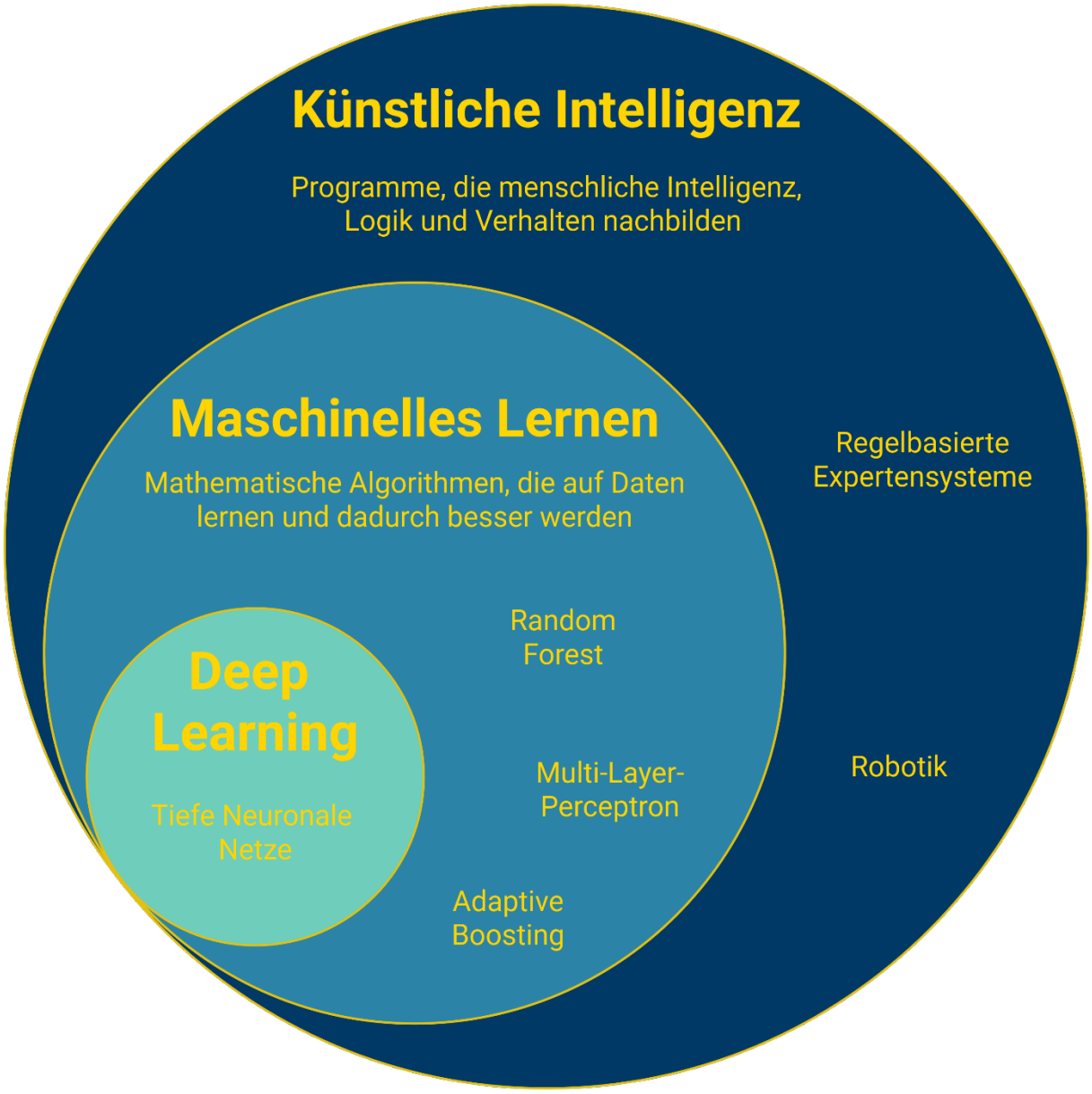

Abbildung 2-1. Taxonomie der Künstlichen Intelligenz, des Maschinellen Lernens und des „Deep Learnings“ (bearbeitete Version des Originals von https://www.kobold.ai/ml-vs-dl/)

Trotz der vielversprechenden Anwendungen ist die Integration von KI und ML in der Medizin mit einigen Herausforderungen verbunden. Der Umgang mit sensiblen medizinischen Daten erfordert strenge Sicherheitsmaßnahmen zum Schutz vor Verstößen und zur Einhaltung von Vorschriften, wie dem HIPAA (Health Insurance Portability and Accountability Act) oder der EU-Datenschutzgrundverordnung

(GDPR). KI-Modelle müssen so konzipiert sein, dass sie ethisch vertretbare und nachvollziehbare Entscheidungen treffen können, insbesondere in Szenarien mit hohem Risiko wie bei der Patientenversorgung (Jobin et al., 2019; Mittelstadt, 2019). Wenn KI-Modelle an der klinischen Entscheidungsfindung beteiligt sind, ergeben sich auch rechtliche Konsequenzen hinsichtlich der Verantwortlichkeit. ML-Modelle können Vorurteile aufrechterhalten oder sogar verstärken, insbesondere wenn sie mit nicht repräsentativen Daten oder retrospektiven Sammlungen gekennzeichneter Proben trainiert und mit anderen oder externen Daten getestet werden (Ferreira et al., 2021; Yu & Kohane, 2019). Die Integration von KI-Modellen in bestehende Gesundheitssysteme ist mit technischen Herausforderungen verbunden und erfordert die Koordinierung zwischen verschiedenen Akteuren. Die Gewährleistung der Interoperabilität dieser Modelle mit verschiedenen Systemen und Technologien ist für ihre effektive Nutzung von entscheidender Bedeutung (Lehne et al., 2019).

Während bei der Entwicklung und Anwendung von KI in der Medizin bisher ein Schwerpunkt auf der Verbesserung der Vorhersagen lag, gibt es heute weitere wichtige Aspekte, wie die Benutzerfreundlichkeit, Erklärbarkeit und Sicherheit von KI-Modellen (Kumar et al., 2023), um die Akzeptanz in der medizinischen Praxis zu steigern. Es ist zu erwarten, dass KI-basierte Technologien mit zunehmender Reife zu einem Standardbestandteil der medizinischen Ausbildung und Praxis werden und zu einer proaktiveren, vorausschauenden und personalisierten Gesundheitsversorgung führen. Dafür wird zunehmend u.a. in der Ausbildung von medizinischen Fachkräften in „Data Literacy“, also die Kompetenz im Umgang mit Daten, investiert (Katzensteiner et al., 2022; S. Kuhn et al., 2018). Führungskräfte und KI-Verantwortliche werden im AI Act verpflichtet, sich ausreichende Kompetenzen im KI-Bereich anzueignen.

Zusammenfassend lässt sich sagen, dass KI und ML die Medizin verändern und Möglichkeiten bieten, Aspekte der Patientenversorgung zu verbessern. Damit diese Technologien ihr volles Potenzial entfalten können, ist es jedoch unerlässlich, die damit verbundenen Herausforderungen pflichtbewusst anzugehen. Auf diese Weise wird sichergestellt, dass KI und ML die Gesundheitsversorgung verbessern und gleichzeitig ethische Standards eingehalten und die Patientenrechte respektiert werden.

Kapitel 2.1 befasst sich mit den Arten des Lernens von KI-Methoden, mit Fokus auf überwachtes Lernen, welches für die Analysen angewendet wurde. Nachdem in Kapitel 2.2 Regressionsverfahren beschrieben werden, beleuchten Kapitel 2.3 und 2.4 ausgewählte Modelle des maschinellen Lernens (ML), die häufig zur Analyse von tabellarisch strukturierten Daten wie z. B. Routinedaten verwendet werden. Kapitel 2.5 gibt einen Einblick in Validierungsmethoden für KI-Methoden. Es bietet auch Strategien zur Verbesserung der Vorhersagekraft von ML-Modellen, vor allem beim Umgang mit stark unausgeglichenen Daten, was im Gesundheitswesen häufig der Fall ist (Kapitel 2.6). Anschließend wird der Aspekt der Erklärbarkeit und Interpretierbarkeit ausführlich beschrieben, der nicht nur für die Erhöhung der Akzeptanz von Entscheidungen auf Basis von ML-Ansätzen, sondern auch für die Einhaltung sich entwickelnder gesetzlicher Rechte für Patienten und andere Nutzer (AI-Act) von entscheidender Bedeutung ist (Kapitel 2.7).

2.1 Arten des Lernens von KI

2.1.1 Überwachtes Lernen (Supervised Learning)

Im Verlauf dieses Weißbuchs wird ein besonderer Fokus auf das überwachte Lernen gelegt, weshalb dieser Ansatz nun zuerst vorgestellt werden soll. Das Ziel des überwachten Lernens besteht darin, auf Basis von Informationen, sogenannten Variablen oder Merkmalen der Stichproben, ein Vorhersagemodell zu entwickeln, welches eine vorher bestimmte Zielvariable vorhersagt. Dabei lernt das Vorhersagemodell im Rahmen des Trainingsprozesses Entscheidungsmuster auf Basis dieser Merkmale. Im mathematischen Sinne entspricht jedes Merkmal dabei einer Dimension, welche gemeinsam den sogenannten Merkmalsraum (Feature Space) aufspannen. Die Gesamtheit aller Merkmale oder Daten für jede Stichprobe wird Datensatz genannt. Im Gegensatz zum Unüberwachten Lernen (Unsupervised Learning) wird beim Überwachten Lernen, je nach Fragestellung, ein Zielmerkmal oder eine Zielvariable innerhalb dieses Merkmalsraums ausgewählt, welches das Ziel der Vorhersage sein wird. Merkmalsräume oder Datensätze, in denen eine Zielvariable enthalten ist, werden im Kontext des maschinellen Lernens häufig annotierte Daten genannt. Annotierte Daten sind zwingend erforderlich, um überwachtes Lernen anwenden zu können. Der restliche Merkmalsraum steht für das ML-Modell zur Verfügung, um Entscheidungsmuster abzuleiten. Der Merkmalsraum kann zusätzlich manuell durch menschliche Auswahl oder mithilfe weiterer Verfahren eingegrenzt werden. Überwachtes Lernen kommt in der Regel zum Einsatz, wenn eine Vorhersage im Sinne einer prädiktiven Schätzung einer kontinuierlichen Variable oder eine Kategorisierung im Sinne einer Klassifikation durch das ML-Modell erbracht werden soll. In einem Datensatz, der zu einer Stichprobe von kardiologischen Patientenfällen entsprechende klinische Merkmale beinhaltet, wäre ein mögliches Vorhersageziel, ob ein Patient einen Myokardinfarkt haben wird oder nicht. Hierfür ist zwingend erforderlich, dass für jeden Patienten die Information vorliegt, ob ein Myokardinfarkt vorlag oder nicht. Diese vereinfachte Darstellung berücksichtigt an dieser Stelle nicht, dass der Datensatz in der Regel vorverarbeitet werden muss, siehe Kapitel 4. Für das Training des Modells wird in der Regel nur eine Untermenge des gesamten Datensatzes verwendet, die im Allgemeinen als Trainingsdaten bezeichnet werden. Die übrigen Daten werden nach dem Training des Modells zur Testung verwendet, um die Vorhersagequalität des Modells zu evaluieren. Das Ziel dieses skizzierten Gesamtprozesses ist es, ein generalisierbares Modell zu entwickeln, welches in der Lage ist, auch auf ungesehenen Daten eine präzise Vorhersage durchzuführen. Kapitel 2.5 erläutert die Details der verschiedenen Evaluierungsverfahren.

2.1.2 Unüberwachtes Lernen (Unsupervised Learning)

Unüberwachtes Lernen wird ebenfalls in vielen ML-Projekten eingesetzt, erfordert jedoch im Gegensatz zum überwachten Lernen keine annotierten Daten. Betrachten wir erneut einen Datensatz von Patienten. Auch hier liegt ein umfassender Merkmalsraum vor. Dieser Raum ermöglicht es dem ML-Verfahren, Assoziationen und Ähnlichkeiten zwischen den Merkmalen selbst zu erkennen und auf dieser Grundlage die vorhandenen Daten in Gruppen zu unterteilen („Clustern“). Auf diese Weise lernt das Modell, dass es Ähnlichkeiten in bestimmten Patientenfällen gibt und kann autonom Untergruppen identifizieren und zuordnen, wie beispielsweise nicht bekannte Subgruppen von Erkrankungen. Auch hier kann der Merkmalsraum angepasst werden. Die Stärke des unüberwachten Lernens liegt jedoch in der Identifizierung unbekannter Zusammenhänge zwischen den Daten, weshalb die Einschränkung des Merkmalsraums als Begrenzung der Lernfähigkeit betrachtet werden kann. Es ist jedoch möglich, auf Basis der Ergebnisse des unüberwachten Lernens den Merkmalsraum zu reduzieren, um die Rechenleistung für nachfolgendes überwachtes Lernen zu minimieren (Dimensionality Reduction). Die Kombination von Verfahren ist also möglich. Zusammenfassend steht beim unüberwachten Lernen eher ein explorativer Charakter im Vordergrund, da neue Zusammenhänge in den Daten entdeckt werden. Dies unterscheidet das Verfahren deutlich von den Aufgabenstellungen des gezielten Vorhersagens im überwachten Lernen.

2.1.3 Bestärkendes Lernen (Reinforcement Learning)

Das dritte Lernkonzept orientiert sich am operanten Konditionieren, einem Prinzip aus der behavioristischen Lernpsychologie, bei dem ein gewünschtes Verhalten (z. B. bei Tieren) mithilfe eines Belohnungs- und Bestrafungssystems trainiert wird. Um dieses Konzept besser im Zusammenhang mit maschinellem Lernen zu verstehen, betrachten wir ein Vorhersagemodell, das die Aufgabe hat, ein Spielzeugauto durch ein Labyrinth zu führen. Über mehrere Trainingszyklen hinweg soll das Modell das Spielzeugauto dazu befähigen, den optimalen Weg durch das Labyrinth zu finden. Während dieser Trainingszyklen gibt das Modell dem Spielzeugauto sowohl korrekte als auch inkorrekte Anweisungen, was zu einer Variation der zurückgelegten Distanz führt. Die Grundidee besteht darin, die Verringerung der Distanz als positives Verhalten des Modells zu interpretieren und eine Zunahme der Distanz als Misserfolg zu bewerten. Diese übergeordnete Bewertung des Modellverhaltens führt zu einer mathematischen Belohnung oder Bestrafung, die sich in der Praxis durch eine Anpassung der Merkmalsgewichte im Modell manifestiert. Das Beispiel illustriert eine praktische Anwendung aus dem Bereich der Robotik, die jedoch auch im Gesundheitswesen in abgewandelter Form Anwendung finden kann, beispielsweise zur Optimierung von Medikamentendosierungen. Zusammenfassend dient das bestärkende Lernen dazu, die Leistung eines ML-Modells anhand einer übergeordneten Verhaltensbewertung zu optimieren.

Im Kontext dieses Weißbuchs liegt der Schwerpunkt auf Ansätzen des überwachten Lernens. Diese Ansätze werden in späteren Kapiteln in Verbindung mit den in Kapitel 1 vorgestellten Routinedaten genutzt, um verschiedene Zielmerkmale vorherzusagen.

2.2 Regressionsverfahren

Regressionsverfahren werden verwendet, um Zusammenhänge zwischen einer abhängigen und einer oder mehreren unabhängigen Variablen zu modellieren. Während die lineare Regression darauf abzielt, kontinuierliche Zielvariablen durch eine lineare Funktion vorherzusagen, ermöglicht die logistische Regression die Prädiktion von Wahrscheinlichkeiten und Gruppenzugehörigkeiten, auch Klassifikation genannt.

2.2.1 Lineare Regression

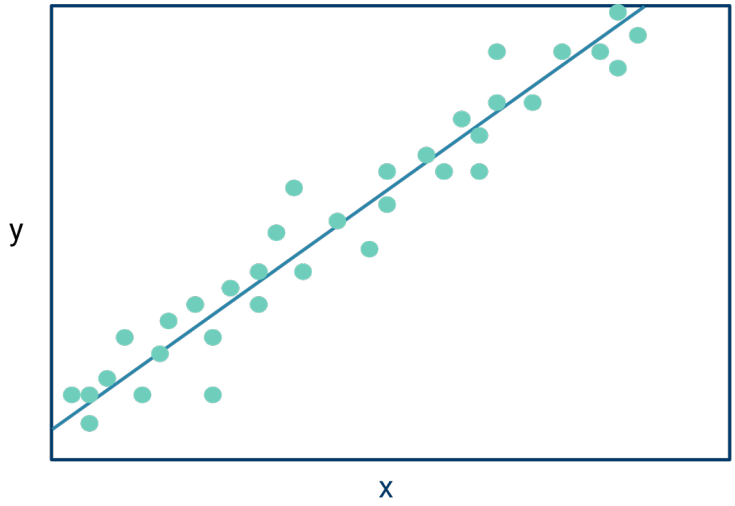

Abbildung 2-2. Lineare Regressionsgerade (blau)

Das Konzept der linearen Regression ist eine statistische Methode, welche auch häufig in Machine Learning verwendet wird. Diese besteht darin, eine lineare Gleichung zu finden, die die Beziehung zwischen den Variablen am besten beschreibt. Diese Gleichung hat die Form:

$$Y = a + bX + \varepsilon$$

Dabei steht Y für die vorhergesagte oder abhängige Variable, X ist die unabhängige Variable, a ist der y-Achsenabschnitt (intercept), b ist der Steigungskoeffizient (slope) und ε ist der Fehlerterm (error), der die unerklärte Variation berücksichtigt. Ziel ist es, die optimalen Werte von a und b zu bestimmen, die die Summe der quadrierten Fehler minimieren, so dass die Linie der besten Anpassung, die die Datenpunkte so gut wie möglich approximiert, wie in Abbildung 2-2 zu erkennen ist.

Die lineare Regression ist ein grundlegendes statistisches Verfahren, mit dem die Beziehung zwischen zwei oder mehr Variablen modelliert wird, wobei der Schwerpunkt auf der Vorhersage einer stetigen Variablen (der abhängigen oder Zielvariablen) auf der Grundlage der Werte einer oder mehrerer anderer Variablen (der unabhängigen oder Prädiktorvariablen) liegt. Sie ist ein wichtiges Instrument im Bereich des maschinellen Lernens und der Statistik und wird häufig für Aufgaben wie Vorhersagen, Trendanalysen und das Verständnis des Zusammenhangs zwischen Variablen eingesetzt.

Die lineare Regression ist vielseitig und umfasst Varianten wie die einfache lineare Regression (mit einem Prädiktor) und die multiple lineare Regression (mit mehreren Prädiktoren). Sie bietet wertvolle Einblicke in Datenmuster, ermöglicht Vorhersagen und ist eine Grundlage für komplexere Algorithmen des maschinellen Lernens. Trotz ihrer Einfachheit ist die lineare Regression nach wie vor ein leistungsfähiges Instrument für eine Vielzahl von Anwendungen in der Praxis.

Für eine ausführliche Einleitung zu dieser Methode verweisen wir auf Urban und Mayerl (Urban & Mayerl, 2008).

Implementierungsbibliotheken: z. B. in Python, Sklearn: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

Erklärbarkeit
Die Slopes in einem linearen Modell können wie folgt interpretiert werden:

- Einfache lineare Regression:

  Positive Steigung: Wenn die unabhängige Variable zunimmt, nimmt die abhängige Variable tendenziell zu.
  Negative Steigung: Wenn die unabhängige Variable zunimmt, nimmt die abhängige Variable tendenziell ab.

- Multiple lineare Regression:

  Positive Steigung für eine unabhängige Variable: Wenn diese Variable zunimmt, während die anderen Variablen konstant bleiben, nimmt die abhängige Variable tendenziell zu.
  Negative Steigung für eine unabhängige Variable: Wenn diese Variable zunimmt und die anderen Variablen konstant bleiben, nimmt die abhängige Variable tendenziell ab.

In beiden Fällen zeigt die Größe der Steigung die Stärke der Beziehung zwischen den Variablen an, und statistische Tests helfen bei der Bewertung ihrer Signifikanz.

2.2.2 Logistische Regression

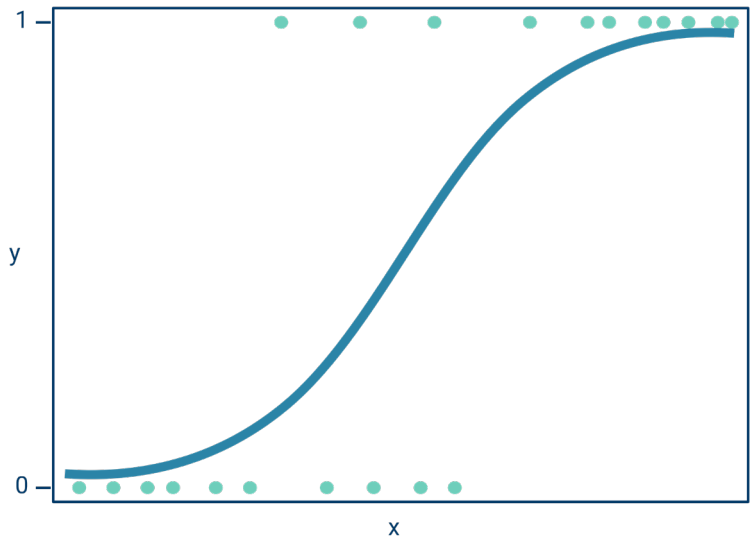

Abbildung 2-3. Logistische Regressionskurve (blau)

Die logistische Funktion, oft auch als Sigmoidfunktion bezeichnet, ist ein wesentlicher Bestandteil der logistischen Regression (siehe Abbildung 2-3). Sie nimmt eine beliebige reelle Zahl und wandelt sie in einen Wert zwischen 0 und 1 um, der als Wahrscheinlichkeit interpretiert werden kann:

$$P(Y = 1) = {}^{1}/_{1 + e^{-(a+b_1x_1+b_2x_2+\ldots+b_nx_n)}}$$

Dabei steht P(Y = 1) für die Wahrscheinlichkeit der positiven Klasse (Klasse 1), und x1, x2, ..., xn sind die unabhängigen Variablen. Zu den Parametern des Modells gehören der Intercept-Term „a“ und die Koeffizienten b1, b2, ..., bn, die die Beziehung zwischen den unabhängigen Variablen und der Wahrscheinlichkeit der Zugehörigkeit zur positiven Klasse bestimmen.

Die logistische Regression ist ein statistisches und maschinelles Lernverfahren, das für binäre und mehrklassige Klassifizierungsaufgaben verwendet wird. Im Gegensatz zur linearen Regression, die kontinuierliche numerische Werte vorhersagt, sagt die logistische Regression die Wahrscheinlichkeit voraus, dass ein Ergebnis zu einer von zwei oder mehr Kategorien gehört. Sie wird häufig in verschiedenen Bereichen eingesetzt, bei denen es um das Auftreten einer bestimmten Ausprägung eines dichotomen Merkmals geht (z. B. Tod ja/nein, Raucher ja/nein).

Die logistische Regression ist ein leistungsstarkes und interpretierbares Werkzeug für Klassifizierungsprobleme, das fundierte Entscheidungen und die Modellierung komplexer Beziehungen zwischen Variablen in einer Vielzahl von Anwendungen ermöglicht. Beispiele wären die Vorhersage des bildgebend bestätigten Wiederauftretens von Prostatakrebs (J. M. Beinecke et al., 2022) und die Analyse der Beziehungen zwischen den Konstitutionstypen der traditionellen chinesischen Medizin und Übergewicht oder Fettleibigkeit (Zhu et al., 2010).

Für eine ausführliche Einleitung zu dieser Methode verweisen wir auf Urban und Mayerl (Urban & Mayerl, 2008).

Implementierungsbibliotheken: z. B. in Python, Sklearn: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html#sklearn.linear_model.LogisticRegression

Erklärbarkeit

Die Interpretation der Koeffizienten ist bei der logistischen Regression von entscheidender Bedeutung:

- Ein positiver Koeffizient ($b_i > 0$) für eine unabhängige Variable x_i bedeutet, dass ein Anstieg des Wertes von x_i zu einer höheren Wahrscheinlichkeit der positiven Klasse führt.
- Ein negativer Koeffizient ($b_i < 0$) bedeutet, dass ein Anstieg des Wertes von x_i die Wahrscheinlichkeit der positiven Klasse verringert.

Darüber hinaus spiegelt die Größe der Koeffizienten die Stärke des Einflusses der Variablen auf die Wahrscheinlichkeit wider. Größere Koeffizienten haben eine stärkere Auswirkung, während kleinere einen schwächeren Effekt haben. Allerdings kann dieser Zusammenhang nicht (wie bei der linearen Regression) linear interpretiert werden. Stattdessen werden die Koeffizienten nach mathematischer Transformation als Chancenverhältnis (Odds Ratio, z.T. auch als exp(B) oder eβ bezeichnet) interpretiert. Die Odds Ratio gibt an, wieviel höher die Chance ist, zur positiven Klasse zu gehören, wenn der Prädiktor um eine Einheit steigt (und alle anderen Prädiktoren gleich bleiben).

2.3 Baumbasierte ML-Verfahren

Baumbasierte ML-Verfahren erweitern den Anwendungsbereich der traditionellen Statistik erheblich, indem sie leistungsfähige Algorithmen zur Analyse und Interpretation großer, komplexer Datensätze einsetzen, die über herkömmliche statistische Methoden hinausgehen. Der Unterschied zwischen den klassischen Methoden und baumbasierten ML-Verfahren liegt im Prozess des Lernens, wobei bei letzteren iterativ die Merkmale nach ihrer Vorhersagequalität selektiert werden. Sie zeichnen sich durch ihre Fähigkeit aus, nichtlineare Zusammenhänge abzubilden und mit hochdimensionalen Daten umzugehen. Zudem ermöglichen sie durch ensemblebasierte Ansätze eine robuste Generalisierung. Die Kombination statistischer Prinzipien mit baumbasierten Lernverfahren erlaubt es, verwertbare Erkenntnisse zu gewinnen und präzise prädiktive Analysen durchzuführen.

2.3.1 Entscheidungsbäume (Decision Trees)

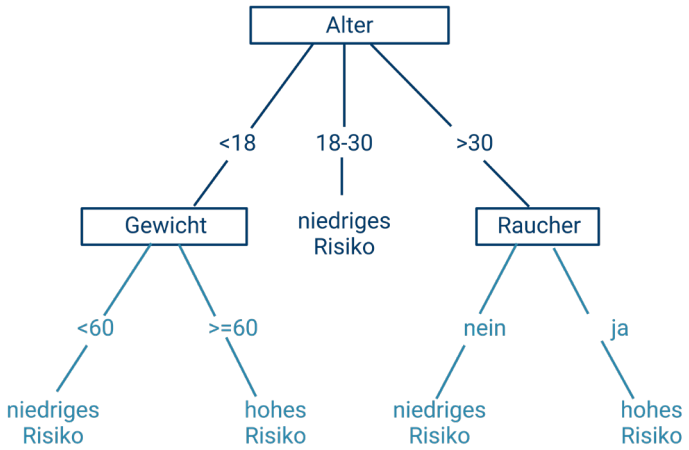

Abbildung 2-4. Beispielhafte Darstellung eines Entscheidungsbaum zur Herzinfarktrisikoabschätzung in Abhängigkeit vom Alter (Angepasst. Quelle: https://www.datacamp.com/tutorial/decision-tree-classification-python)

Wie Abbildung 24 dargestellt, sind Entscheidungsbäume wie Flussdiagramme strukturiert, die eine baumartige Hierarchie von Knoten aufweisen. Zu diesen Knoten gehören der Wurzelknoten, der die anfängliche Entscheidung oder Frage auf der Grundlage eines Merkmals darstellt, interne Knoten, die nachfolgende Fragen zu verschiedenen Merkmalen stellen, und Blattknoten, die die Endpunkte des Baums bilden und die endgültigen Vorhersagen oder Ergebnisse liefern.

Während des Trainingsprozesses teilen die Entscheidungsbäume den Datensatz an jedem internen Knoten in Teilmengen auf. Das Ziel ist es, möglichst homogene Teilmengen in Bezug auf die Zielvariable zu erstellen. Dies geschieht durch die Auswahl des Merkmals und ggf. eines Schwellenwerts, der die Unreinheit oder den Fehler bei einem bestimmten Teilungsschritt am stärksten minimiert. Gängige Maße für die Unreinheit sind die Gini-Impurity und Entropie. Der Prozess der Aufteilung der Daten in Teilmengen auf der Grundlage von Merkmalen erfolgt rekursiv. Interne Knoten teilen die Daten so lange auf, bis ein Abbruchkriterium erfüllt ist, z. B. das Erreichen einer bestimmten Tiefe, das Erreichen von reinen Klassenknoten (alle Daten in einem Knoten gehören zur selben Klasse) oder wenn keine weitere Verbesserung der Unreinheit erzielt werden kann.

Entscheidungsbäume stellen eine hierarchische Struktur dar, die einen fließdiagrammartigen Entscheidungsprozess nachahmt. Entscheidungsbäume werden besonders für ihre Einfachheit, Interpretierbarkeit und Effektivität bei der Erfassung komplexer Beziehungen in Daten geschätzt. Sie können jedoch zu einer Überanpassung (Overfitting) neigen, insbesondere wenn der Baum zu tief wird. Um dies abzumildern, werden häufig Techniken wie das Beschneiden und die Begrenzung der Tiefe des Baums angewendet.

Entscheidungsbäume bilden eine Grundlage für komplexere Ensemble-Methoden wie Random Forests und Gradient Boosting und sind nach wie vor ein wertvolles Instrument für die Untersuchung von Datenmustern, die Erstellung von Vorhersagen und die Gewinnung von Erkenntnissen in verschiedenen Bereichen, vom Gesundheitswesen (Speiser et al., 2018; Starr et al., 2020) über das Finanzwesen bis hin zu Empfehlungssystemen.

Für eine ausführliche Einleitung zu dieser Methode verweisen wir auf Müller et al. (Müller et al., 2017).

Implementierungsbibliotheken: z. B. in Python, Sklearn: 1.10. Decision Trees — scikit-learn 1.3.0 documentation.

Erklärbarkeit

Entscheidungsbäume sind in hohem Maße interpretierbar, was sie zu einem wertvollen Hilfsmittel macht, um die Gründe für Vorhersagen zu erläutern. Sie können den Pfad vom Wurzelknoten zu einem Blattknoten zumindest grundsätzlich leicht nachvollziehen, um zu verstehen, wie Entscheidungen getroffen wurden. Allerdings können auch Entscheidungsbäume mit vielen Merkmalen sehr komplex werden, insbesondere sofern eine entsprechende Größe des Baumes bzw. eine große Anzahl an Entscheidungsebenen berücksichtigt wurde.

2.3.2 Random Forest

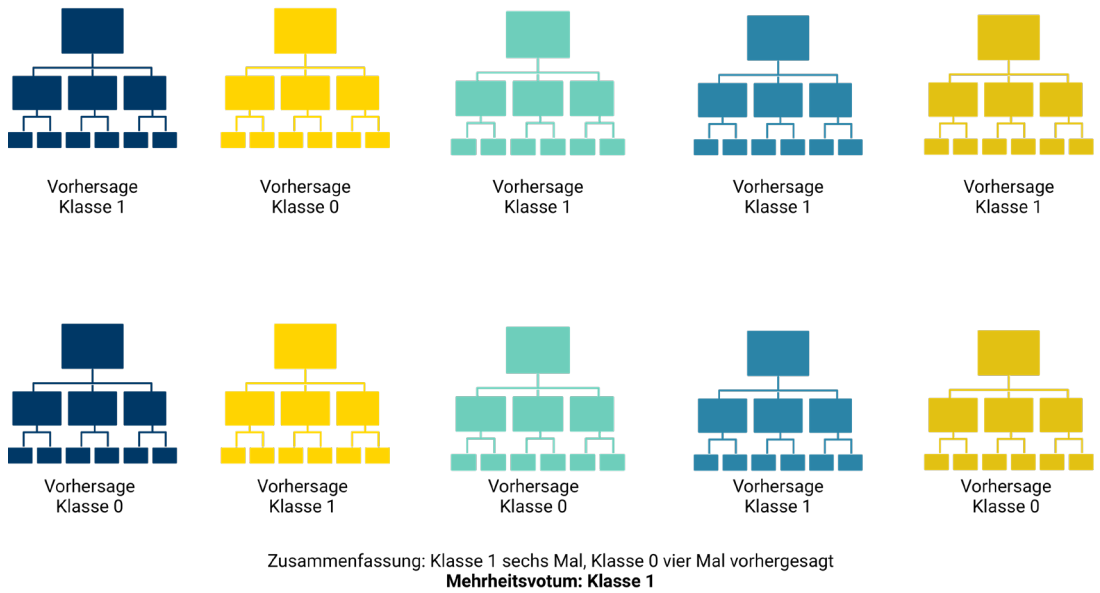

Abbildung 2-5. Darstellung eines Random Forests bestehend aus mehreren Entscheidungsbäumen

Der Prozess beginnt mit der zufälligen Auswahl von Teilmengen der Trainingsdaten (Bootstrapping) und Merkmalen (Unterraum der Merkmale) für jeden Entscheidungsbaum. Jeder dieser Entscheidungsbäume im Random Forest wird auf einer anderen Teilmenge trainiert, was zu mehr Vielfalt führt und das Risiko von Overfitting verringert.

Bei der Vorhersage erstellt jeder Baum im Random Forest eine Vorhersage (siehe Abbildung 2-5), und bei Klassifizierungsaufgaben wird der Modus (häufigste Klasse) dieser Vorhersagen als endgültige Ensemble-Vorhersage verwendet. Bei der Regression wird der Durchschnitt der Vorhersagen der einzelnen Bäume gebildet.

Random Forests (RFs) sind eine leistungsstarke Ensemble Learning-Methode, die für ihre Robustheit und Vielseitigkeit bei der Lösung von Klassifizierungs- und Regressionsproblemen bekannt ist. Ensemble-Modelle sind Techniken, die die Vorhersagen von mehreren Einzelmodellen kombinieren, um die Gesamtleistung und Robustheit zu verbessern. Diese Modelle machen sich die Vielfalt der Basismodelle zunutze, um deren individuelle Schwächen zu umgehen, was zu einer verbesserten Genauigkeit und Verallgemeinerung führt. Durch die Nutzung der Stärken verschiedener Modelle erreichen Ensemble-Methoden oft eine bessere Vorhersageleistung als ein einzelnes Modell. Bei dieser Ensemble-Methode werden während des Trainings mehrere Entscheidungsbäume (siehe Abschnitt 2.2.3) erstellt und dann ihre Vorhersagen kombiniert, um die Genauigkeit zu verbessern und eine Überanpassung (Overfitting) zu vermeiden.

RFs finden in verschiedenen Bereichen Anwendung. Ihre Anpassungsfähigkeit, Robustheit und Fähigkeit, komplexe Datensätze zu verarbeiten, machen RFs zu einer beliebten Wahl für eine Vielzahl von Aufgaben des maschinellen Lernens, wie z. B. die Vorhersage der Sterblichkeit und des Krankenhausaufenthalts bei Herzversagen (Angraal et al., 2020) oder die Schätzung des Flüssigkeitsflusses in gekrümmten Rohren (Narayanan et al., 2021).

Für eine ausführliche Einleitung zu dieser Methode verweisen wir auf Müller et al. (Müller et al., 2017).

Implementierungsbibliotheken: z. B. in Python, Sklearn: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Erklärbarkeit

Aufgrund ihrer Komplexität und Größe werden RFs oft als Black-Box-Modelle eingestuft, obwohl sie Wichtigkeitsmetriken (Feature Importance Scores) wie die Gini Importance liefern. Die Metrik für die Feature Importance in RFs ist ein wichtiges Instrument, um den Einfluss der einzelnen Variablen zu den Vorhersagen des Modells zu verstehen. Es hilft bei der Identifizierung der Merkmale, die den größten Einfluss auf Vorhersagen haben. Die Gini Importance misst beispielsweise, wie oft ein Merkmal zur Aufteilung der Daten in allen Entscheidungsbäumen im RF verwendet wird, und wird als Summe der Abnahme der Gini-Impurity (ein Maß für die Datenunreinheit) für jede Aufteilung im Baum, in dem das Merkmal verwendet wird, berechnet. Merkmale, die zu einer größeren Verringerung der Gini-Impurity führen, werden als wichtiger angesehen. Es gibt noch andere Feature Importance Metriken, wie die Mean Decrease in Accuracy (MDA), Mean Decrease in Node Impurity (MDNI) und die Permutation Importance (Hastie et al., 2009).

Neben der Feature Importance können sogenannte modellagnostische post-hoc Methoden (siehe Abschnitt 2.5.1) wie Shapley Values (Lundberg & Lee, 2017) verwendet werden, nachdem das Modell trainiert wurde, um statistisch zu berechnen, welche Eingangsmerkmale das Ergebnis des RF beeinflusst haben. Konkret kann hier Tree SHAP (Lundberg et al., 2020) verwendet werden, welches eine schnelle und genaue Methode zur Schätzung von Shapley Values für baumbasierte Modelle darstellt.

2.3.3 Adaptive Boosting

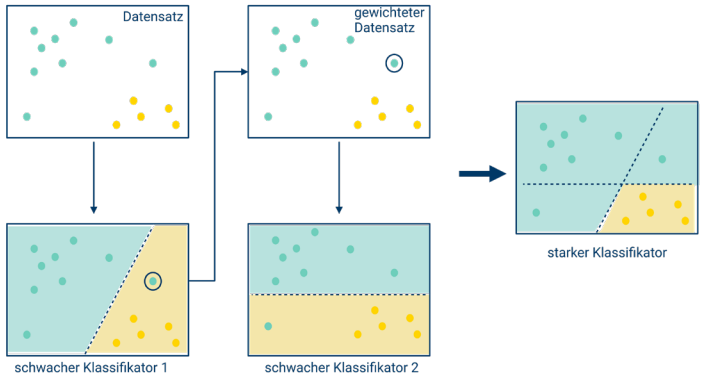

Abbildung 2-6. Beispielhafte Darstellung des Adaptive Boosting Algorithmus

AdaBoost, kurz für Adaptive Boosting, ist eine Ensemble Learning-Methode, die für ihre Fähigkeit bekannt ist, die Genauigkeit schwacher Lerner zu verbessern und robuste, leistungsstarke Modelle zu erstellen. Es funktioniert, indem nacheinander eine Reihe schwacher Modelle trainiert wird - Modelle, die etwas besser sind als Raten - und falsch klassifizierten Datenpunkten in jeder nachfolgenden Runde mehr Gewicht verliehen wird. AdaBoost findet in verschiedenen Bereichen Anwendung, darunter Krankheitsvorhersage, Gesichtserkennung und Textklassifizierung, wo präzise und anpassungsfähige Modelle erforderlich sind. So wurde es beispielsweise zur Vorhersage von Hirntumortypen auf MR-Bilddaten (Minz & Mahobiya, 2017) oder zur Klassifizierung der Sterblichkeit von Sepsispatienten (Peng et al., 2022) eingesetzt.

Zu Beginn weist der AdaBoost-Algorithmus allen Datenpunkten im Trainingssatz die gleiche Gewichtung zu. Danach wird für jede Iteration ein schwaches Modell ausgewählt, das die gewichteten Klassifizierungsfehler minimiert. Dieses schwache Modell wird dann auf den Daten mit gewichteten Stichproben trainiert, und das Gewicht für die Vorhersage des schwachen Modells wird auf der Grundlage seiner Fehlerquote berechnet. Nach dem Training werden die Gewichte der Datenpunkte aktualisiert, wobei zuvor falsch klassifizierte Datenpunkte mehr Bedeutung bekommen. Schließlich werden die Vorhersagen aller schwachen Modelle zu einer gewichteten Summe kombiniert, um das endgültige Modell zu bilden, wie es auch in Abbildung 2-6 dargestellt ist.

Für eine ausführliche Einleitung zu dieser Methode verweisen wir auf Schapire (Schapire, 2013).

Implementierungsbibliotheken: z. B. in Python, Sklearn: https://scikit-learn.org/stable/modules/ensemble.html#AdaBoost

Erklärbarkeit
Aufgrund der Komplexität des endgültigen Modells, das aus mehreren schwachen Modellen besteht, können wir nicht nachvollziehen, welche Merkmale das Ergebnis beeinflusst haben. Post-hoc-Methoden wie die Shapley-Werte (Lundberg & Lee, 2017) können nach dem Training des Modells verwendet werden, um statistisch zu berechnen, welche Eingangsmerkmale das Ergebnis des endgültigen Modells beeinflusst haben.

2.4 Künstliche Neuronale Netze

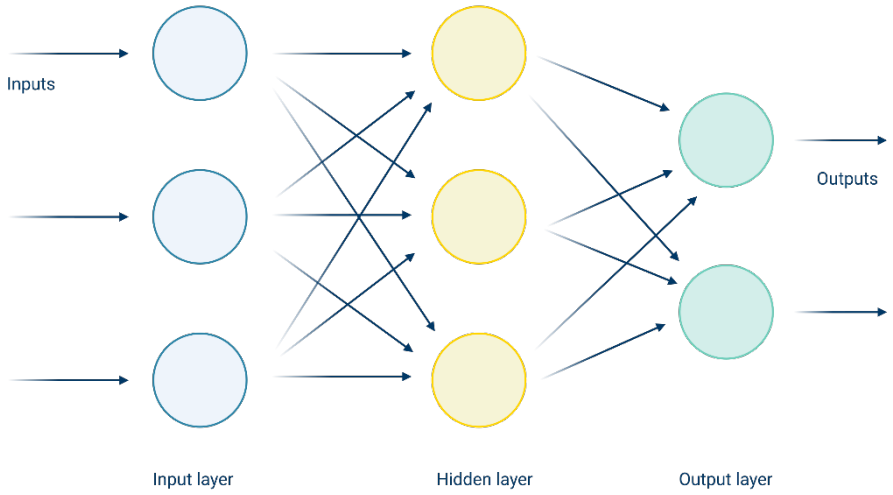

Abbildung 2-7. Darstellung eines einfachen Künstliches Neuronalen Netzes, mit einer Eingangsschicht, versteckten Schicht und Ausgangsschicht

Künstliche Neuronale Netze (KNNs) sind aus miteinander verbundenen Schichten (layers) aufgebaut, die die Eingabe-, die versteckte und die Ausgabeschicht umfassen (siehe Abbildung 2-7). Diese Schichten verarbeiten Daten durch gewichtete Verbindungen und Aktivierungsfunktionen (Activation Functions) und ermöglichen es KNNs, komplizierte Muster und Beziehungen zwischen verschiedenen Datentypen zu erkennen. Während der Trainingsphase passen KNNs die Gewichte und Biases der Neuronen dynamisch an und nutzen dabei einen Backpropagation-Algorithmus und Optimierungstechniken, um Vorhersagefehler zu minimieren. KNNs erfordern für jede spezifische Aufgabe die Formulierung einer maßgeschneiderten Verlustfunktion (Loss Function), um den Lernprozess der KNN-Modelle zu steuern. So dienen beispielsweise Mean Squared Error und Cross-Entropy Verlustfunktionen als Zielfunktionen für einfache Regressions- und Klassifikationsaufgaben. Für detaillierte Beschreibungen der Algorithmen verweisen wir auf (Calin, 2020).

Künstliche Neuronale Netze (KNN) ist ein allgemeiner Begriff für alle maschinellen Lernalgorithmen, die von der Struktur und Funktion biologischer neuronaler Netze inspiriert sind. Nach dem Vorbild der Struktur und Funktion des menschlichen Gehirns sind KNNs Rechenmodelle, die in der Lage sind, komplexe Muster zu lernen, Vorhersagen zu treffen und eine Vielzahl von Aufgaben zu lösen. KNNs sind darauf ausgelegt, beliebig komplexe Beziehungen zwischen hochdimensionalen Eingaben und Ausgaben zu approximieren. Sie umfassen eine Vielzahl von Architekturen, darunter Single-Layer Perceptrons, Multi-Layer Perceptrons (MLPs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Transformer, Generative Modelle und Diffusionsmodelle. KNNs werden häufig für maschinelle Lernaufgaben verwendet, die von einfacher Regression und Klassifizierung bis hin zu Objekterkennung, Segmentierung und Detektion reichen.

In der heutigen Zeit spielen KNNs eine zentrale Rolle bei der Entwicklung robuster und zuverlässiger Algorithmen, nicht nur in der Informatik, sondern auch in verschiedenen wissenschaftlichen Bereichen wie Biologie, Klimawissenschaft, Medizin und Ozeanografie. Im Jahr 2021 erzielte DeepMind mit seinem KI Modell AlphaFold einen Durchbruch, indem es Algorithmen auf der Grundlage von CNNs nutzte, um eine der wichtigsten Herausforderungen der Biologie zu lösen: die Vorhersage der 3D-Struktur von Proteinen (Jumper et al., 2021). Weitere Anwendungen von KNNs sind die Schätzung der Insulinwirkung nach einer Mahlzeit (Mosquera-Lopez et al., 2023) sowie die Klassifizierung von Untertypen der Alzheimer-Krankheit (Shigemizu et al., 2023).

Für eine ausführliche Einleitung zu dieser Methode verweisen wir auf Müller et al. (Müller et al., 2017).

Implementierungsbibliotheken: z. B. in Python: https://scikit-learn.org/stable/modules/neural_networks_supervised.html , https://pytorch.org/ , https://www.tensorflow.org/

Erklärbarkeit

KNNs sind von Natur aus Black-Box-Methoden, d. h. man kann nicht nachvollziehen, welche Variablen das Ergebnis beeinflusst haben. Post-hoc Methoden wie Shapley Values (Lundberg & Lee, 2017) können nach dem Training des Modells verwendet werden, um statistisch zu berechnen, welche Eingangsmerkmale das Ergebnis des KNN beeinflusst haben. Es gibt viele andere Erklärungsmethoden, die zum Beispiel in der Captum-Bibliothek (https://captum.ai/, Kokhlikyan et al., 2020) implementiert sind.

2.5 Validierung und Optimierung von KI-Methoden

Die Validierung von KI-Methoden zielt darauf ab, die Qualität oder Leistung eines maschinellen Lernmodells zu quantifizieren, insbesondere um zu bestimmen, wie robust und zuverlässig sich ein trainiertes maschinelles Lernmodell mit neuen oder unbekannten Daten verhält. Mit anderen Worten: Die Vorhersagekapazität, die während der Trainingsphase mithilfe spezifischer Metriken gemessen wird, unterscheidet sich nicht wesentlich, wenn diese für neue, zum Training nicht verwendete Daten (z. B. Test- und Validierungsdatensätze) ermittelt werden. Die Bewertung und Validierung eines Modells sind von entscheidender Bedeutung, um die Leistung des ML-Modells sicherzustellen, nachdem es für den Einsatz in realen Szenarien bereitgestellt wird.

Die Leistung eines Modells wird anhand spezifischer Metriken und „scoring“ bestimmt. Diese werden in Kapitel 3 behandelt. Zur Bewertung bei Klassifizierungsproblemen wird üblicherweise die Genauigkeit (Accuracy) als Bewertungsparameter verwendet. Metrische Funktionen wie die Konfusionsmatrix, Precision-Recall, und Receiver Operating Characteristic-Kurve (ROC-Kurve) werden ebenfalls eingesetzt. Bei Regressionsproblemen wird unter anderem der mittlere quadratische Fehler (mean squared error) als Bewertungsparameter benutzt. Der Wert, den ein Modell basierend auf der verwendeten Metrik erhält, wird als Score bezeichnet. Bei „scoring“-Metriken liegt der Bewertungswert zwischen null und eins. Ein Wert nahe eins weist eine sehr gute Vorhersagekraft des Modells auf, ein Wert nahe null weist eine sehr geringe Vorhersagekraft auf. Im Fall von metrischen Funktionen, wie ROC oder der Konfusionsmatrix, werden funktionsspezifische Werte und graphische Profile verwendet, um sie zu bewerten und zu vergleichen. Die Evaluierungsmetrik eines Modells ist mit der Verlustfunktion verknüpft, die vom Optimierungsalgorithmus verwendet wird, um die Parameter des Modells in aufeinanderfolgenden Iterationen anzupassen. Das Ziel ist hier, die Leistung des Modells in jeder Iteration zu verbessern. Der Score wird erst am Ende des Trainings aus dem Testdatensatz ermittelt. Die spezifischen Rechenvorschriften der Scores und Verlustfunktionen werden als Metriken bezeichnet.

In sklearn sind verschiedene scoring- und metrische Funktionen implementiert, siehe: https://scikit-learn.org/stable/modules/model_evaluation.html#classification-metrics

2.5.1 Train-Test-Split

Die Train-Test-Split-Methode ist eine häufig verwendete Validierungsmethode im Bereich des maschinellen Lernens. Hierbei wird der vorhandene Datensatz in zwei bzw. drei Teile aufgeteilt: Trainingsdaten und Testdaten. Wenn Parameter-Optimierung durchgeführt wird, wird ein zusätzliches Split in Training und Validierungsdaten verwendet. Die Trainingsdaten werden für das eigentliche Training des Modells verwendet, indem das Modell die zugrundeliegenden Muster in den Daten lernen kann. Die Validierungsdaten werden dazu verwendet, die Leistung des Modells während des Trainings zu bewerten, um die Hyperparameter abzustimmen und eine Überanpassung zu verhindern. Die Testdaten werden letztendlich genutzt, um die Endleistung des Modells zu bewerten. Ein Vorteil dieser Methode liegt in ihrer einfachen Anwendung. Allerdings ist die Methode auch von der zufälligen Datenaufteilung abhängig, was insbesondere bei Daten mit einer geringen Anzahl an Beobachtungen zu Variabilität führen kann. Zudem besteht die Gefahr von Overfitting, wenn das Modell zu stark auf die Validierungsdaten angepasst wird, sodass der Test-Fehler bei hoher Modellkomplexität weiter ansteigt (siehe Abbildung 2-8).

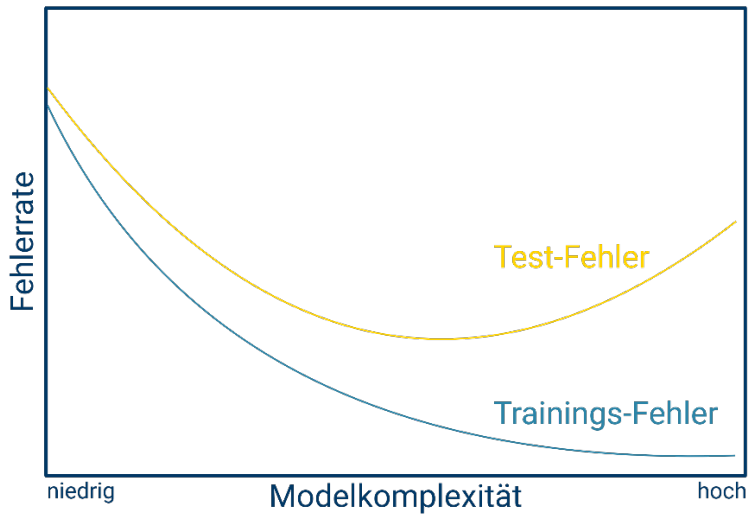

Abbildung 2-8. Konzeptionelle Darstellung von Test-Fehler und Trainings-Fehler bei steigender Modelkomplexität

2.5.2 Kreuzvalidierung

Die Kreuzvalidierung ist eine Erweiterung der Train-Test Split-Methode, die darauf abzielt, die Robustheit eines Modells zu überprüfen, insbesondere bei kleinen Datensätzen.

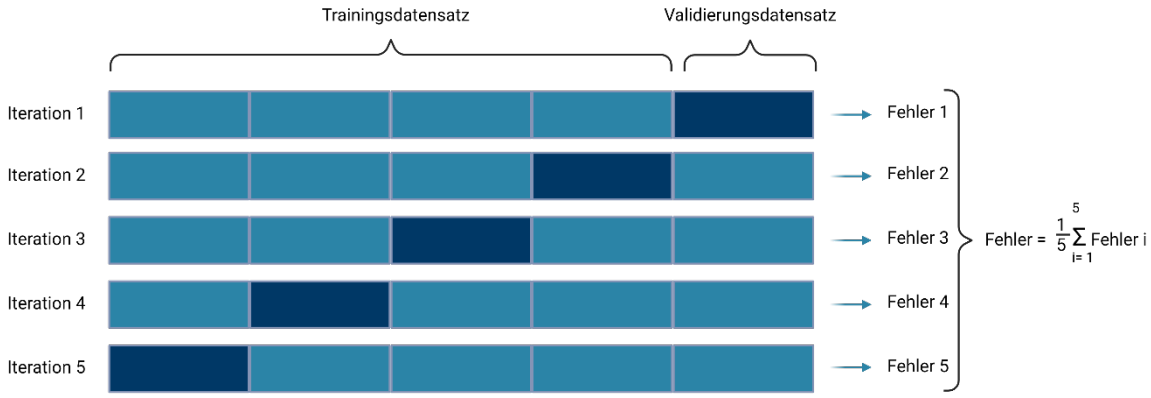

Abbildung 2-9. Darstellung einer 5-fold Kreuzvalidierung

Eine häufig verwendete Methode ist die k-fold Kreuzvalidierung, bei der der Datensatz in k Teile aufgeteilt wird (James et al., 2013), wie es in Abbildung 2-9 dargestellt ist. Das Modell wird dann k-mal trainiert und validiert, wobei in jedem Durchgang ein anderer Teil als Validierungsdatensatz verwendet wird. Demgegenüber ist die leave-one-out Kreuzvalidierung eine spezielle Form, bei der jeder Datenpunkt einmal als Validierungsset fungiert. Dies ist besonders nützlich bei sehr kleinen Datensätzen, da sie maximalen Nutzen aus den vorhandenen Daten zieht. Der Vorteil der Kreuzvalidierung liegt darin, den Einfluss der zufälligen Datenaufteilung zu reduzieren und somit eine zuverlässige Schätzung der Modellleistung zu bieten. Allerdings geht dies mit einem höheren Rechenaufwand einher und bei ungleichmäßig verteilten Daten können Probleme wie unterrepräsentierte Klassen in einem Teil der folds, Overfitting auf folds, oder falsche Einschätzung der Generalisierbarkeit, auftreten, was insbesondere bei kleineren Datenbeständen ein Problem sein kann. Um eine proportionale Verteilung der Klassen

zu erreichen, kann eine stratifizierte Kreuzvalidierung verwendet werden. Bei stratifizierter Kreuzvalidierung werden für jeden k-fold oder jede Iteration die gleichen prozentualen Anteile der Klassen verwendet, wie die im gesamten Datensatz vorhanden sind, d. h. bevor der Train-Test-Split des Datensatzes kreuzvalidiert wird. Generell kann die Robustheit durch Wiederholung des Kreuzvalidierungsprozesses erhöht werden; hierfür kann die sklearn-Funktion „repeated cross-validation“ verwendet werden. Die Gittersuche (grid search) oder Zufallssuche (random search) werden speziell zur Optimierung der Modell-Hyperparameter verwendet. Gittersuche kombiniert jeden der gegebenen Werte der Hyperparameter, um sie in verschiedenen ML-Modellen zu testen; Zufallssuche kombiniert sie zufällig, wobei einige mögliche Kombinationen nicht getestet werden. Gittersuche wird zusammen mit Kreuzvalidierung verwendet, und die Funktion GridSearchCV wird dafür z. B. in sklearn verwendet.

2.5.3 Monte-Carlo-Validierung

Die Monte Carlo Validierung (Xu & Liang, 2001) basiert auf dem Prinzip der zufälligen Stichprobenziehung und eignet sich gut bei sehr großen Datensätzen. Hierbei werden mehrere zufällige Teilmengen des Datensatzes erstellt, um die Modellleistung zu bewerten (siehe Abbildung 2-10). Diese Methode ermöglicht es, eine breite Palette von Datenkombinationen zu berücksichtigen und eine statistisch robuste Bewertung der Modellleistung zu erhalten.

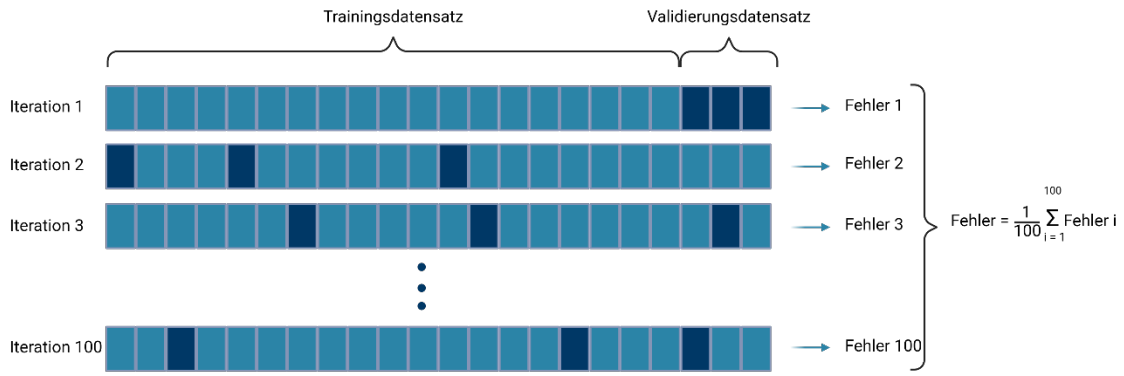

Abbildung 2-10. Darstellung einer Monte Carlo Validierung mit 100 zufälligen Splits

Ein Vorteil der Monte-Carlo-Validierung liegt in ihrer Anpassungsfähigkeit an große Datensätze, bei denen es schwierig ist, alle möglichen Kombinationen zu berücksichtigen. Allerdings geht dies mit einem höheren Rechenaufwand einher, da der Prozess wiederholt durchgeführt werden muss. Zudem können die Ergebnisse aufgrund der zufälligen Stichprobenziehung variieren. Die Monte-Carlo-Validierung hängt stark von der Verteilung der Stichproben ab, die die Varianz beeinflusst (Kalos & Whitlock, 2008; Metropolis & Ulam, 1949). Monte Carlo weist beispielsweise im Vergleich zur Kreuzvalidierung eine geringe Varianz auf. Die Ergebnisse sind wiederholbarer als bei der Kreuzvalidierung, jedoch auf Kosten einer höheren Verzerrung (Bias).

Insgesamt sind diese Validierungsmethoden entscheidend, um zu überprüfen, ob ML-Methoden zuverlässig und generalisierbar in verschiedenen Anwendungsbereichen funktionieren.

2.6 Umgang mit unbalancierten Daten

Unbalancierte Daten sind ein häufiges Problem beim maschinellen Lernen (ML), das auftritt, wenn die Anzahl der Beobachtungen in jeder Klasse innerhalb eines Datensatzes nicht ungefähr gleich ist. Dieses Ungleichgewicht kann die Leistung von ML-Modellen erheblich beeinträchtigen und zu verzerrten oder ungenauen Ergebnissen führen. Viele reale Datensätze sind von Natur aus unbalanciert. Bei der Erkennung von Betrug beispielsweise ist die Zahl der betrügerischen Transaktionen in der Regel viel geringer als die Zahl der rechtmäßigen Transaktionen. Ebenso können in der medizinischen Diagnostik Krankheiten im Vergleich zu gesunden Fällen selten sein. Bei unausgewogenen Daten neigen viele ML-Modelle dazu, die Mehrheitsklasse zu bevorzugen und die Minderheitsklasse, die häufig von größerem Interesse ist, zu ignorieren oder falsch zu klassifizieren. Dies kann nur zum Teil durch eine geeignete Wahl der Optimierungsparameter behoben werden, siehe 2.4.2.

2.6.1 Upsampling und Downsampling

Up- und Downsampling von Daten sind Techniken für den Umgang mit unbalancierten Daten. Downsampling ist ein einfacher Ansatz, bei dem Stichproben aus der größeren Klasse vor dem Training entfernt werden. Dadurch können die ML-Modelle auf einem balancierten Datensatz trainiert werden. Dieser Ansatz ist in der Regel nur durchführbar, wenn große Datenmengen verfügbar sind.

Upsampling ist eine Technik, bei der neue synthetische Stichproben der kleineren Klasse algorithmisch generiert werden. Zu diesen Methoden gehören SMOTE, ADASYN, Gaussian-Noise-Upsampling und weitere (Beinecke & Heider, 2021). Der SMOTE-Algorithmus (Chawla et al., 2002) wurde für numerische Daten entwickelt und generiert neue Datenpunkte für die kleinere Klasse basierend auf den k-nächsten-Nachbarn jedes Datenpunktes. Dieser Ansatz ist jedoch nicht direkt auf nominale und kategoriale Daten anwendbar. SMOTE-N modifiziert den ursprünglichen SMOTE-Algorithmus, um nominale und kategoriale Daten zu verarbeiten, indem er die Ähnlichkeit zwischen den Kategorien bei der Erzeugung synthetischer Stichproben berücksichtigt. Dazu können Techniken wie die Hamming-Distanz verwendet werden, die die Anzahl der Attribute zählt, bei denen sich zwei Instanzen unterscheiden, oder anspruchsvollere Methoden, die die Verteilung der Kategorien im Datensatz berücksichtigen.

2.6.2 Gewichtung der Fehler-Funktion

Ein weiterer Ansatz ist das Benutzen einer gewichteten Fehler-Funktion (auch Loss-Funktion) beim Training des ML-Modells (überwachtes Lernen, siehe Kapitel 2.1.1). Hierbei wird die Fehler-Funktion als Ausgangspunkt gewählt, um das Modell zu optimieren. Ziel beim Training ist es, die Fehler-Funktion zu minimieren, wodurch das Modell besser wird. Damit das ML-Modell besser darin wird, die kleinere Klasse vorherzusagen, ist es naheliegend das Modell stärker zu „bestrafen“, wenn es diese falsch vorhersagt. Dies kann man umsetzen, indem man in der Fehler-Funktion Fehler in der Klassifizierung der kleineren Klasse mit einem großen Gewicht vergrößert und Fehler in der Klassifizierung der größeren Klasse mit einem kleineren Gewicht verringert. Dadurch wird es während des Trainings für das ML-Modell wichtiger die Elemente der kleineren Klasse richtig zu klassifizieren.

2.7 Interpretierbare und erklärbare KI im Gesundheitswesen

Interpretierbare und erklärbare KI (XAI) ist in der Medizin von entscheidender Bedeutung, da sie sicherstellt, dass Entscheidungen von KI-Modellen transparent, verständlich und vertrauenswürdig sind, was die Einhaltung ethischer und regulatorischer Standards ermöglicht (Jobin et al., 2019; Mittelstadt, 2019). Entscheidungen im Gesundheitswesen sind von Natur aus komplex und können schwerwiegende Auswirkungen auf die Patientenversorgung haben. Medizinische Fachkräfte müssen oft die Grundlage von Empfehlungen der ML-Modelle verstehen, um sie effektiv in ihre Entscheidungsprozesse zu integrieren. XAI bietet Einblicke in die Art und Weise, wie KI-Modelle ihre Ergebnisse generieren, und hilft Klinikern sicherzustellen, dass diese Vorschläge fundiert und vertretbar sind (Chaddad et al., 2023). Ein Mangel an Vertrauen in KI-Modelle ist immer noch eine der größten Blockaden, KI als klinische Entscheidungsunterstützungssysteme zu implementieren. Patienten und ihre Gesundheitsdienstleister müssen den Instrumenten und Methoden vertrauen, die in den Behandlungsplänen eingesetzt werden. Die Erklärbarkeit von KI trägt dazu bei, dieses Vertrauen aufzubauen, indem sie die Technologie sowohl für Ärzte/Ärztinnen als auch für Patienten/Patientinnen transparenter und nachvollziehbarer macht. Wenn ein Patient versteht, wie ein KI-Modell zu einer Diagnose gekommen ist oder warum eine bestimmte Behandlung empfohlen wurde, wird sein Vertrauen in die Behandlung steigen.

Die Medizin ist stark reguliert, um das Wohlergehen der Patienten zu schützen und sicherzustellen, dass die Behandlungen nach ethischen Grundsätzen durchgeführt werden. Die Erklärbarkeit von KI-Modellen trägt dazu bei, diese gesetzlichen Anforderungen zu erfüllen, indem sie Prüfpfade und Nachweise für dessen Entscheidungen liefert. Dies ist besonders wichtig für die Einhaltung von Gesetzen wie der EU-Datenschutzgrundverordnung (GDPR) (*Data Protection in the EU - European Commission*, 2023) oder dem EU AI Act (*AI Act | Shaping Europe's Digital Future*, 2024; *European AI Office | Shaping Europe's Digital Future*, o. J.), die das Recht auf eine Erklärung automatisierter Entscheidungen vorsieht.

Zusammenfassend lässt sich sagen, dass XAI in der Medizin wichtige Anforderungen in Bezug auf Sicherheit, Vertrauen, Verständnis und Einhaltung von Vorschriften erfüllt. XAI ermöglicht eine effektivere Integration von KI in die Gesundheitsversorgung, indem sie fortschrittliche Technologie mit den differenzierten Anforderungen der medizinischen Praxis in Einklang bringt und letztlich darauf abzielt, die Ergebnisse für die Patienten zu verbessern und gleichzeitig hohe Standards in Bezug auf Gesundheitsversorgung und Ethik zu wahren.

2.7.1 Taxonomie

In diesem Abschnitt führen wir Klassifikationssysteme (Taxonomien) von interpretierbarer und erklärbarer KI ein. Sowohl interpretierbare als auch erklärbare KI zielen darauf ab, die Transparenz und das Verständnis zu verbessern, aber sie unterscheiden sich im Detailgrad und in der Tiefe der Erklärungen, die sie liefern.

Interpretierbare KI-Modelle können Erklärungen über ihre Funktionsweise und Entscheidungsprozesse liefern, ohne dass zusätzliche Hilfsmittel erforderlich sind. Interpretierbare KI ermöglicht es, das „Wie" und „Warum" von Entscheidungen durch die Struktur und Parameter des KI-Modells direkt zu erfassen. So ist beispielsweise ein lineares Regressionsmodell (siehe Kapitel 2.2) von Natur aus interpretierbar, weil man sehen kann, wie die Eingabemerkmale gewichtet werden, um eine Ausgabe zu erzeugen. In ähnlicher Weise bieten Entscheidungsbäume (siehe Kapitel 2.3) einen klaren Pfad von Entscheidungen, was sie leicht interpretierbar macht. Ein wichtiger Aspekt interpretierbarer KI ist, dass die Erklärung meistens eine direkte Folge der Einfachheit und Transparenz des Modells ist.

Bei der erklärbaren KI hingegen geht es oft um Techniken, die zur Erklärung des Verhaltens eines Modells verwendet werden, unabhängig davon, ob das Modell selbst interpretierbar ist. Dies ist besonders wichtig für komplexe Modelle wie Neuronale Netze (siehe Kapitel 2.4) oder große Random Forests (siehe Kapitel 2.3.2), bei denen der Entscheidungsprozess nicht einfach zu verstehen ist. Erklärbare KI verwendet häufig Techniken, um Erklärungen für diese Modelle abzuleiten und Einblicke in die Entstehung der Modellvorhersage zu geben, nachdem das Modell trainiert wurde. Solche Techniken werden eingesetzt, um herauszufinden, welche Merkmale für bestimmte Entscheidungen wichtig waren, und helfen so, das Verhalten des Modells zu verstehen.

Die Unterscheidung zwischen interpretierbarer und erklärbarer KI ist in Bereichen von entscheidender Bedeutung, in denen das Verständnis der Gründe für eine Entscheidung ebenso wichtig ist wie die Entscheidung selbst, z. B. im Gesundheitswesen, im Finanzwesen und bei rechtlichen Anwendungen. Im Folgenden werden wir noch auf weitere Kategorisierungen von interpretierbarer und erklärbarer KI eingehen, sowie konkrete XAI-Modelle. Für weitere Literatur zu Taxonomien für erklärbare und interpretierbare KI verweisen wir auf (Arrieta et al., 2020; Hanif et al., 2023; Schneeberger et al., 2023; Speith, 2022).

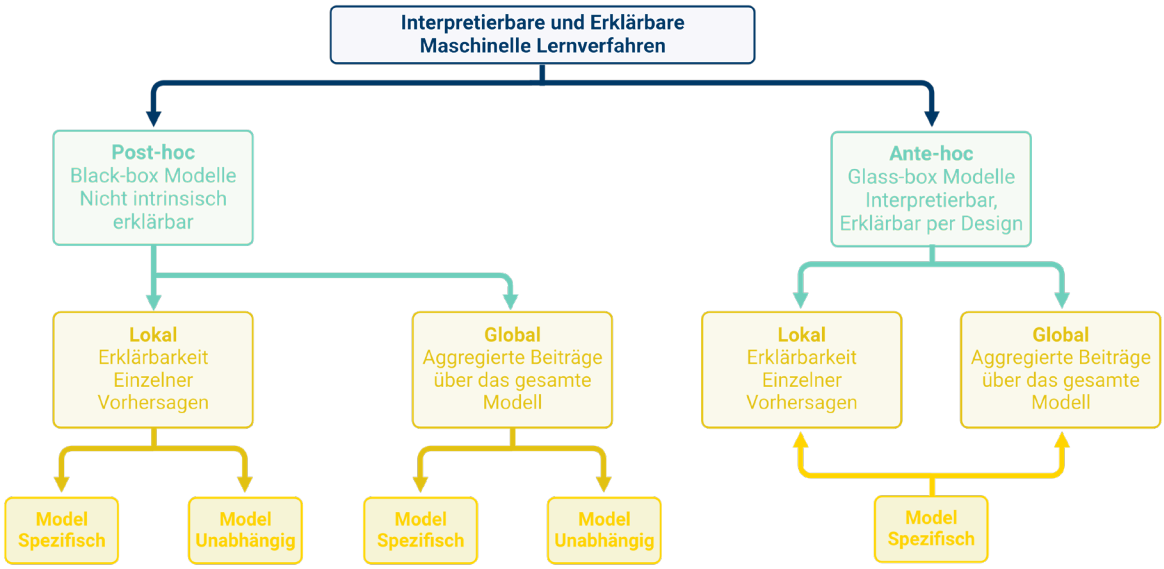

Abbildung 2-11. Taxonomie der interpretierbaren und erklärbaren KI-Lernverfahren

Post-hoc vs. Ante-hoc

Die Kategorisierung von interpretierbarer und erklärbarer KI wird häufig auch als „Ante-hoc“ und „Post-hoc“ bezeichnet. Diese Begriffe beschreiben, wann die Erklärungen in Bezug auf den Modellbildungsprozess generiert werden. Eine Darstellung dieser Taxonomie ist in Abbildung 2-11 zu sehen.

„Ante-hoc" KI-Modelle beziehen sich dabei auf Modelle, die per Definition interpretierbar sind, wie die Lineare Regression oder Entscheidungsbäume. Diese werden auch häufig als „Glass-box“ oder „White-box“ Modelle bezeichnet. Bei diesen Modellen ist der Erklärungsprozess im Modell selbst integriert.

„Post-hoc“ Modelle sind XAI-Modelle, welche genutzt werden, um KI-Modelle, die sogenannte „Black-box“ Modelle sind (nicht interpretierbar), zu erklären. Hierbei handelt es sich um XAI-Modelle, die auf ein KI-Modell angewendet werden, nachdem dies eine Entscheidung getroffen hat. Diese Modelle erzeugen also im Nachhinein eine Erklärung für die Entscheidung.

Lokal vs. Global

Eine weitere Unterteilung von XAI-Modellen bezieht sich auf die Reichweite der Erklärungen. Lokale XAI-Modelle erzeugen Erklärungen für einzelne Vorhersagen eines KI-Modells. Ihr Fokus liegt darauf, Einsichten in den Entscheidungsprozess für eine spezifische Instanz/Datenpunkt zu liefern. Globale XAI-Modelle hingegen versuchen den Entscheidungsprozess eines KI-Modells über alle Instanzen/Datenpunkte zu erklären. Diese Modelle versuchen die allgemeine Logik, Regeln und Muster eines KI-Modells zu durchleuchten, anstatt sich auf nur eine individuelle Vorhersage zu fokussieren.

Modellspezifisch vs. Modellunabhängig

Des Weiteren können XAI-Modelle noch darüber unterschieden werden, ob sie nur auf ein oder mehrere KI-Modelle angewandt werden können. Modellspezifische XAI-Modelle sind meist für einen speziellen Typ von KI-Modell entwickelt und nutzen zum Beispiel die Modellarchitektur oder innere Logik des Modells, um Erklärungen zu erzeugen. Durch die Integration der Informationen über das KI-Modell können modellspezifische XAI-Modelle häufig detailliertere und genauere Erklärungen über den Entscheidungsprozess des KI-Modells liefern. Modellunabhängige XAI-Modelle sind, wie ihr Name schon sagt, nicht für ein spezielles KI-Modell entwickelt. Sie wurden entwickelt, um auf jegliches KI-Modell angewendet zu werden.

2.7.2 Lokale post-hoc Erklärbare KI-Modelle

Im Rahmen unserer Analysen von Neuronalen Netzen haben wir drei unterschiedliche Arten von lokalen post-hoc XAI-Modellen betrachtet. Auf diese verschiedenen Arten gehen wir hier noch einmal genauer ein.

Perturbationsbasierte Methoden

Perturbationsbasierte KI-Methoden verändern die Daten sukzessive und beobachten, wie sich diese Änderungen auf die Vorhersagekraft des Klassifikationsmodells auswirken. Im Falle von Bilddaten erfolgt diese zum Beispiel durch Löschung einzelner Pixel im Bild oder im Falle von tabularen Daten, durch Entfernen einzelner Variablen. Andere Methoden wie Shapley Values (Shapley, 1952) basieren auf kooperativer Spieltheorie und versuchen, das Ergebnis eines Spiels (Vorhersage des Klassifikationsmodells) fair auf alle Spieler (Variablen) zu verteilen. Shapley-Werte werden als der durchschnittliche marginale Beitrag eines Merkmalswertes über alle möglichen Merkmalsuntergruppen hinweg. Da dies in der praktischen Umsetzung eine zu hohe Rechenleistung erfordert, wurde SHAP (Lundberg & Lee, 2017) eingeführt, welches als eine Näherung an die tatsächlichen Shapley Values zu verstehen ist. Ein Beispiel für die Visualisierung von lokalen SHAP-Werten ist in Abbildung 2-12 aufgeführt, welche einen Wasserfallplot einiger Eingabevariablen des Heart Disease Datensatzes (Janosi et al., 1988) darstellt. SHAP ist eine Modell-unabhängige XAI-Methode.

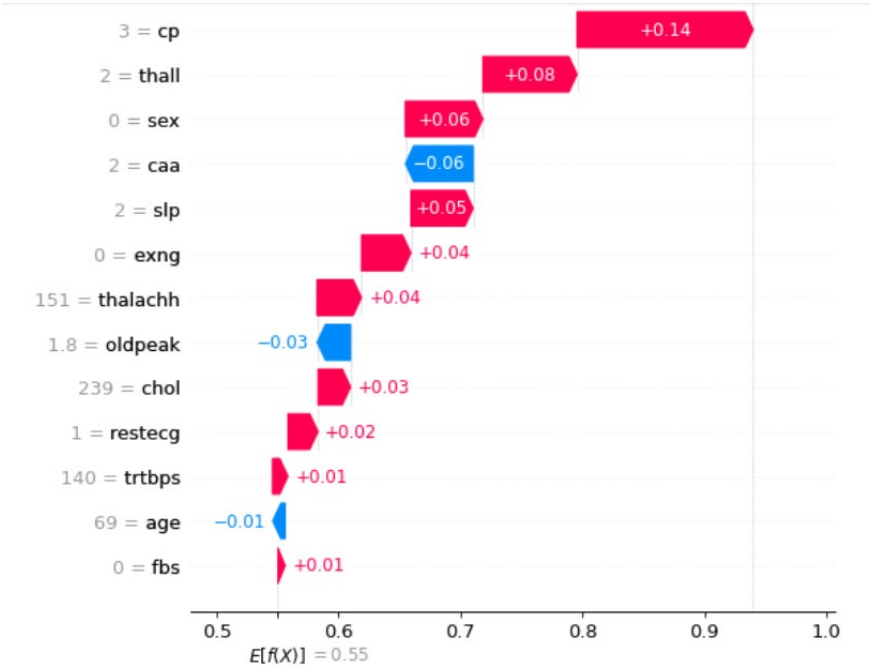

Abbildung 2-12. Wasserfallplot für lokale SHAP-Werte. Links sind die Eingabevariablen mit ihren Ausprägungen gegeben und rechts ihre SHAP-Werte in einem Wasserfallplot dargestellt. Der Output des Neuronalen Netzes für die Eingabe (Erwartungswert) ist ebenfalls in den Plot eingezeichnet. Das Beispiel wurde auf dem Heart Disease Datensatz (Janosi et al., 1988) erzeugt.

Gradienten-basierte Methoden

Gradienten-basierte KI-Methoden nutzen den Backpropagation-Algorithmus eines Neuronales Netzes, um den Gradienten des Netzes bezüglich der Eingabevariablen durch das Neuronale Netz zu propagieren. Dies ermöglicht es ihnen, Attributionen durch das gesamte Neuronale Netzwerk zu propagieren, um schließlich Attributionen für jede Eingabe Variable zu erhalten. Damit handelt es sich hierbei um Modell-abhängige XAI-Methoden, da sie den Backpropagation-Algorithmus eines Neuronalen Netzes ausnutzen.

Die Integrated Gradients (Sundararajan et al., 2017) Methode berechnet den mittleren Gradienten entlang eines geradlinigen Weges von einem Ausgangspunkt (typischerweise Null) zur Eingabe. Die Autoren rechtfertigen den Ausgangspunkt bei Null mit der Argumentation, dass das Zuweisen von Schuld impliziert, dass etwas fehlt oder nicht existiert.

Surrogate Methoden

Surrogate Methoden benutzen erklärbare KI-Methoden, um nicht erklärbare Methoden zu approximieren. Da dies aufgrund der Komplexität der nicht erklärbaren Methoden global nicht möglich ist, erfolgt die Approximation nur in einem kleinen lokalen Bereich um den Datenpunkt, für den eine Erklärung erzeugt werden soll. Die Idee hinter LIME (Ribeiro et al., 2016) ist ein Genauigkeits-Interpretierbarkeits-Gleichgewicht. Das interpretierbare Stellvertretermodell sollte das nicht erklärbare Modell lokal gut genug approximieren, aber gleichzeitig nicht zu komplex werden, um noch erklärbar zu bleiben. In der Praxis wird nur die Genauigkeitsfunktion minimiert und die Komplexität der erklärbaren Methode wird vorher festgelegt.

Quellen

AI Act | Shaping Europe's digital future. (2024, April 23). https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Angraal, S., Mortazavi, B. J., Gupta, A., Khera, R., Ahmad, T., Desai, N. R., Jacoby, D. L., Masoudi, F. A., Spertus, J. A., & Krumholz, H. M. (2020). Machine Learning Prediction of Mortality and Hospitalization in Heart Failure With Preserved Ejection Fraction. *JACC. Heart Failure*, *8*(1), 12–21. https://doi.org/10.1016/j.jchf.2019.06.013

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.

Beinecke, J., & Heider, D. (2021). Gaussian noise up-sampling is better suited than SMOTE and ADASYN for clinical decision making. *BioData Mining*, *14*(1), 49. https://doi.org/10.1186/s13040-021-00283-6

Beinecke, J. M., Anders, P., Schurrat, T., Heider, D., Luster, M., Librizzi, D., & Hauschild, A.-C. (2022). Evaluation of machine learning strategies for imaging confirmed prostate cancer recurrence prediction on electronic health records. *Computers in Biology and Medicine*, *143*, 105263. https://doi.org/10.1016/j.compbiomed.2022.105263

Chaddad, A., Peng, J., Xu, J., & Bouridane, A. (2023). Survey of Explainable AI Techniques in Healthcare. *Sensors (Basel, Switzerland)*, *23*(2), 634. https://doi.org/10.3390/s23020634

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Data protection in the EU - European Commission. (2023, Juli 4). https://commission.europa.eu/law/law-topic/data-protection/data-protection-eu_en

European AI Office | Shaping Europe's digital future. (o. J.). Abgerufen 29. April 2024, von https://digital-strategy.ec.europa.eu/en/policies/ai-office

Ferreira, M. V., Almeida, A., Canario, J. P., Souza, M., Nogueira, T., & Rios, R. (2021). Ethics of AI: Do the Face Detection Models Act with Prejudice? *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Vir-Tual Event*, *Proceedings, Part II 10*, 89–103.

Hanif, A., Beheshti, A., Benatallah, B., Zhang, X., Habiba, F., E., & Shahabikargar, M. (2023). A Com-prehensive Survey of Explainable Artificial Intelligence (XAI) Methods: Exploring Transparency and Inter-pretability. *International Conference on Web Information Systems Engineering*, 915–925.

Hasti, Tibshirani & Friedman (2009). The Elements of Statistical Learning (2nd Edition). Springer-Verlag.

James, G.,Witten, D., Hastie, T., Tisbshirani, R. (2013). *An introduction ot statistical learning* (Bd. 112). Springer.

Jamshidi, A., Pelletier, J. P., & Martel-Pelletier, J. (2019). Machine-learning-based patient-specific prediction models for knee osteoarthritis. *Nature Reviews Rheumatology*, *15*(1), 49–60.

Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1988). *Heart Disease*. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399.

Jumper, J., Evans, R., & Pritzel, A. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kalos, M. H., & Whitlock, P. A. (2008). *Monte Carlo methods* (2., revised and enlarged ed). WILEY-VCH.

Karaarslan, E., & Aydın, D. (2021). An artificial intelligence–based decision support and resource management system for COVID-19 pandemic. In *Data Science for COVID-19* (S. 25–49). Academic Press.

Katzensteiner, M., Vogel, S., Hüsers, J., Richter, J., & Bott, O. J. (2022). Towards a Didactic Concept for Heterogeneous Target Groups in Digital Learning Environments—First Course Implementation. *Journal of Personalized Medicine*, *12*(5), Article 5. https://doi.org/10.3390/jpm12050696

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., & Reblitz-Richardson, O. (2020). *Captum: A unified and generic model interpretability library for pytorch*.

Kuhn, S., Kadioglu, D., Deutsch, K., & Michl, S. (2018). Data Literacy in der Medizin. *Der Onkologe*, *24*(5), 368–377. https://doi.org/10.1007/s00761-018-0344-9

Kumar, Y., Koul, A., Singla, R., & Ijaz, M. F. (2023). Artificial intelligence in disease diagnosis: A systematic literature review, synthesizing framework and future research agenda. *Journal of Ambient Intelligence and Humanized Computing*, *14*(7), 8459–8486. https://doi.org/10.1007/s12652-021-03612-z

Lehne, M., Sass, J., Essenwanger, A., Schepers, J., & Thun, S. (2019). Why digital medicine depends on interoperability. *Npj Digital Medicine*, *2*(1), 1–5. https://doi.org/10.1038/s41746-019-0158-1

Leondes, C. T. (2002). *Expert systems: The technology of knowledge management and decision making for the 21st century*. https://www.sciencedirect.com/book/9780124438804/expert-systems

Lundberg, S. M., Erion, G., & Chen, H. (2020). From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*, *2*, 56–67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17*, 4768–4777.

Mathis-Ullrich, F., & Scheikl, P. M. (2021). Robotik im Operationssaal – (Ko-)Operieren mit Kollege Roboter. *Gastroentero-loge*, *16*, 25–34. https://doi.org/10.1007/s11377-020-00496-x

Metropolis, N., & Ulam, S. (1949). The Monte Carlo Method. *Journal of the American Statistical Association*, *44*(247), 335–341. https://doi.org/10.1080/01621459.1949.10483310

Minz, A., & Mahobiya, C. (2017). MR Image Classification Using AdaBoost for Brain Tumor Type. *2017 IEEE 7th Interna-Tional Advance Computing Conference (IACC*, 701-705,. https://doi.org/10.1109/IACC.2017.0146.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507.

Mosquera-Lopez, C., Wilson, L. M., & El Youssef, J. (2023). Enabling fully automated insulin delivery through meal detection and size estimation using Artificial Intelligence. Npj Digit. *Med*, *6, 39*. https://doi.org/10.1038/s41746-023-00783-1

Müller, A. C., Guido, S., & Rother, K. (2017). *Einführung in Machine Learning mit Python: Praxiswissen Data Science*. O'Reilly. https://books.google.de/books?id=UK72DwAAQBAJ

Narayanan, G., Jain, P., Choudhury, A., Dutta, P., Kalita, K., & Barsocchi, P. (2021). Random Forest Regression-Based Machine Learning Model for Accurate Estimation of Fluid Flow in Curved Pipes. *Processes*, *9*(11), Article 11. https://doi.org/10.3390/pr9112095

Norris, C. M., Ghali, W. A., Saunders, L. D., Brant, R., Galbraith, D., Faris, P., Knudtson, M. L., & APPROACH Investigators. (2006). Ordinal regression model and the linear regression model were superior to the logistic regression models. *Journal of Clinical Epidemiology*, *59*(5), 448–456. https://doi.org/10.1016/j.jclinepi.2005.09.007

Peng, L., Peng, C., Yang, F., Wang, J., Zuo, W., Cheng, C., Mao, Z., Jin, Z., & Li, W. (2022). Machine learning approach for the prediction of 30-day mortality in patients with sepsis-associated encephalopathy. *BMC Medical Research Methodology*, *22*, 183. https://doi.org/10.1186/s12874-022-01664-z

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). „Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Schapire, R. E. (2013). Explaining AdaBoost. In B. Schölkopf, Z. Luo, & V. Vovk (Hrsg.), *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik* (S. 37–52). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-41136-6_5

Schneeberger, D., Röttger, R., Cabitza, F., Campagner, A., Plass, M., Müller, H., & Holzinger, A. (2023). The Tower of Babel in Explainable Artificial Intelligence (XAI). In A. Holzinger, P. Kieseberg, F. Cabitza, A. Campagner, A. M. Tjoa, & E. Weippl (Hrsg.), *Machine Learning and Knowledge Extraction* (S. 65–81). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-40837-3_5

Schünemann, H. J., Dorn, J., Grant, B. J., Winkelstein, W., & Trevisan, M. (2000). Pulmonary function is a long-term predictor of mortality in the general population: 29-year follow-up of the Buffalo Health Study. *Chest*, *118*(3), 656–664. https://doi.org/10.1378/chest.118.3.656

Shapley, L. S. (1952). A Value for n-Person Games. In H. W. Kuhn & A. W. Tucker (Hrsg.), *Contributions to the Theory of Games (AM-28), Volume II* (S. 307–318). Princeton University Press. https://doi.org/10.1515/9781400881970-018

Shigemizu, D., Akiyama, S., & Suganuma, M. (2023). Classification and deep-learning–based prediction of Alzheimer disease subtypes by using genomic data. *Transl Psychiatry*, *13*, 232. https://doi.org/10.1038/s41398-023-02531-1

Speiser, J. L., Karvellas, C. J., Shumilak, G., Sligl, W. I., Mirzanejad, Y., Gurka, D., Kumar, A., & Kumar, A. (2018). Predicting in-hospital mortality in pneumonia-associated septic shock patients using a classification and

regression tree: A nested cohort study. *Journal of Intensive Care*, *6*, 66. https://doi.org/10.1186/s40560-018-0335-3

Speith, T. (2022). A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2239–2250. https://doi.org/10.1145/3531146.3534639

Starr, L. T., Ulrich, C. M., Junker, P., Huang, L., O'Connor, N. R., & Meghani, S. H. (2020). Patient Risk Factor Profiles Associated with the Timing of Goals-of-Care Consultation Before Death: A Classification and Regression Tree Analysis. *The American journal of hospice & palliative care*, *37*(10), 767–778. https://doi.org/10.1177/1049909120934292

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning* (S. 3319–3328). PMLR.

Urban, D., & Mayerl, J. (2008). *Regressionsanalyse: Theorie, Technik und Anwendung*. VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-91194-6

Viros, A., Fridlyand, J., Bauer, J., Lasithiotakis, K., Garbe, C., Pinkel, D., & Bastian, B. C. (2008). Improving Melanoma Classification by Integrating Genetic and Morphologic Features. *PLOS Medicine*, *5*(6), e120. https://doi.org/10.1371/journal.pmed.0050120

Xu, Q. S., & Liang, Y. Z. (2001). Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, *56*(1), 1–11.

Yu, K. H., & Kohane, I. S. (2019). Framing the challenges of artificial intelligence in medicine. *BMJ Quality & Safety*, *28*(3), 238–241.

Zhu, Y., Wang, Q., Wu, C., Pang, G., Zhao, J., Shen, S., Xia, Z., & Yan, X. (2010). [Logistic regression analysis on relationships between traditional Chinese medicine constitutional types and overweight or obesity]. *Zhong xi yi jie he xue bao = Journal of Chinese integrative medicine*, *8*(11), 1023–1028. https://doi.org/10.3736/jcim20101104

3 Gütemaße für Vorhersagemodelle

Vorhersagemodelle liefern in der Regel keine perfekten Vorhersagen, sondern in zumindest einem Teil der Fälle fehlerhafte Vorhersagen. Bei dichotomen Ereignissen können zwei Arten von Fehler auftreten: Eine Kategorie wird vorhergesagt, obwohl die andere vorliegt und umgekehrt. Aus der Häufigkeit dieser Fehler können entsprechende Gütemaße, wie Sensitivität (Recall), Spezifität, positiver Vorhersagewert (Precision) und negativer Vorhersagewert berechnet werden. Auch Maße wie der Matthews-Korrelationskoeffizient (MCC) und der F1-Score basieren auf diesen Kennwerten und kombinieren mehrere Informationen, z. B. aus Sensitivität und positivem Vorhersagewert.
Da Vorhersagemodelle häufig nicht eine dichotome Vorhersage liefern, sondern Eintrittswahrscheinlichkeiten für das dichotome Outcome, muss zunächst ein Schwellenwert festgelegt werden, um die vorhergesagte Wahrscheinlichkeit einer Kategorie zuzuordnen und die entsprechenden Gütemaße zu berechnen. Es gibt aber auch Gütemaße, die alle möglichen Schwellenwerte berücksichtigen. Dazu zählen die Receiver Operating Characteristic (ROC) Kurve und die Precision-Recall (PR) Kurve.

Prädiktionsmodelle, die im Rahmen von sowohl klassisch statistischen als auch ML-basierten Verfahren erstellt werden, liefern in der Regel keine perfekten Vorhersagen. So wird ein Prädiktionsmodell typischerweise – zumindest für einen Teil der Fälle – fehlerhafte Vorhersagen liefern. Die Häufigkeit und das Ausmaß dieser Fehler werden dabei zwischen unterschiedlichen Modellen und Verfahren (z. B. verschiedenen ML-Methoden) variieren. Um die Güte von Prädiktionsmodellen bzw. deren Prognose sinnvoll und objektiv zu vergleichen, sind Kennzahlen erforderlich, die Rückschlüsse auf die Güte einer Vorhersage – möglichst gleichermaßen auch für unterschiedliche Herangehensweisen und Konstellationen – erlauben. Ideal wäre, wenn eine einzige Kennzahl alle relevanten Aspekte der Vorhersagegüte in sich vereint. Nur dann könnten alle Vorhersagemodelle anhand dieser einen Kennzahl hinsichtlich ihrer Güte auch in eine eindeutige Rangfolge von schlechter bis besser gebracht werden.

Nachfolgend werden relevante Kennzahlen zur Beurteilung der Vorhersagegüte erläutert und hinsichtlich ihrer Eignung und Bedeutung bei der Beurteilung von Prädiktionsmodellen diskutiert. Im Rahmen von KI-THRUST wurden ausschließlich Prognosen für das Eintreten von binär/dichotom kodierten Ereignissen betrachtet (z. B. Todesfälle, Wiedereinweisungen). Gütemaße zu Vorhersagen für stetige Outcomes (wie z. B. für die Vorhersage von Behandlungskosten) werden daher an dieser Stelle nicht erläutert.

3.1 Sensitivität und Spezifität, PPV und NPV

Ein Prädiktionsmodell liefert in den Grundzügen vergleichbare Ergebnisse wie ein medizinischer Test. Weder Prädiktionsmodelle noch medizinische Tests liefern dabei in der Praxis perfekte Ergebnisse. So ist beispielsweise bei einem medizinischen Test auf das Vorliegen einer SARS-CoV-2-Infektion stets davon auszugehen, dass der Test nicht jede einzelne Infektion entdeckt. In diesem Fall liefert der Test fälschlich ein als „negativ“ bezeichnetes Ergebnis, obwohl ein Patient real „positiv“ auf das Vorliegen der Infektion hätte getestet werden sollen. Zugleich kann ein Test in der Praxis jedoch auch eine Infektion anzeigen, obwohl diese real nicht vorliegt. Hier wäre das Testergebnis fälschlich „positiv“, obwohl die Person eigentlich ein „negatives“ Ergebnis, also keine Infektion, aufweist. Zur Beurteilung der Güte von entsprechenden Tests spielen vor diesem Hintergrund zwei Kennzahlen eine entscheidende Rolle – die Sensitivität und die Spezifität eines Tests.

- Als **Sensitivität** (englisch: sensitivity, auch recall) eines Tests bezeichnet man den Anteil der positiv getesteten Personen an allen real positiven Personen. Die Sensitivität gibt Auskunft darüber, wie viele der real Erkrankten anteilig durch einen Test als solche erkannt werden.
- Als **Spezifität** (englisch: specificity) eines Tests bezeichnet man den Anteil der negativ getesteten Personen an allen real negativen Personen in einer Population. Die Spezifität gibt Auskunft darüber, wie viele der real Gesunden anteilig durch einen Test als solche erkannt werden.

Beide Kennzahlen bilden gemeinsam traditionell die wohl relevantesten Kennzahlen zur Beurteilung eines medizinischen Tests. Dabei ist offensichtlich, dass keine der beiden Kennzahlen für sich allein genommen zur Beurteilung eines medizinischen Tests geeignet ist:

- Würde ein Test pauschal bei allen Personen das Vorliegen einer Erkrankung anzeigen, würde er damit immer auch alle real positiven Personen „erkennen“ – die Sensitivität dieses Tests wäre immer 100 Prozent – erst der Blick auf die Spezifität mit einem Wert von 0 Prozent würde die Untauglichkeit des Tests offenbaren.
- Umgekehrt könnte ein Test pauschal auch bei keiner Person das Vorliegen einer Erkrankung anzeigen, womit alle real gesunden Personen korrekt als solche klassifiziert werden – die Spezifität des Tests wäre 100 Prozent – erst der Blick auf die Sensitivität mit einem Wert von 0 Prozent würde die Untauglichkeit des Tests offenbaren.

Analoge Aussagen gelten auch für die Ergebnisse zur Sensitivität und Spezifität von Prognosemodellen. Entsprechende Ergebnisse zur Beurteilung der Güte eines medizinischen Tests oder eines Prognosemodells lassen sich in Vier-Felder-Tafeln (engl. Confusion MatrixConfusion Matrix) darstellen. Dabei ist es im Fall des Prognosemodells erforderlich, einen Schwellenwert (engl. Threshold) auf Basis der vorhergesagten Wahrscheinlichkeiten festzulegen, ab der die Ereigniskategorie wechselt (z. B. von gesund zu krank). Die Darstellung einer Vier-Felder-Tafel in Abbildung 3-1 soll zugleich genutzt werden, um weitere Kennzahlen zur Beurteilung der Ergebnisse von Tests oder Prognosemodellen zu erläutern.

		Wahrer Status		
		betroffen (krank)	**nicht betroffen (gesund)**	
Prognose- oder Test-Ergebnis	**positiv**	**A** = 80 richtig positiv (TP)	**B** = 90 falsch positiv (FP)	**A + B** Positive gesamt
	negativ	**C** = 20 falsch negativ (FN)	**D** = 810 richtig negativ (TN)	**C + D** Negative gesamt
		A + C Betroffene gesamt	**B + D** nicht Betroffene gesamt	**A + B + C + D** Gesamtpopulation

in Klammern: TP: true positive, FP: false positive, FN: false negative, TN: true negative

Kennwert	Berechnung	Beispielberechnung
Sensitivität	A / (A + C)	80 / (80 + 20) = 0,8 = 80 %
Spezifität	D / (B + D)	810 / (90 + 810) = 0,9 = 90 %
Positiver Vorhersagewert (PPV)	A / (A + B)	80 / (80 + 90) = 0,47 = 47 %
Negativer Vorhersage-wert (NPV)	D / (C + D)	810 / (20 + 810) = 0,98 = 98 %
Betroffenenanteil, Prävalenz (wahr)	(A + C) / (A + B + C + D)	(80 + 20) / (80 + 90 + 20 + 810) = 0,1 = 10 %

Abbildung 3-1. Vier-Felder-Tafel (engl. Confusion Matrix) zur Beurteilung der Güte von Vorhersagen (oben) und assoziierte Kennwerte (unten)

Neben Sensitivität und Spezifität sind in Abbildung 3-1 auch die Kennzahlen PPV und NPV sowie in der letzten Zeile die Prävalenz beziehungsweise der reale Betroffenenanteil aufgeführt, der im eingefügten Zahlenbeispiel 10 Prozent beträgt. Von 1.000 Personen der Gesamtpopulation waren im genannten Beispiel 100 Personen real erkrankt und hätten von einem idealen Test oder einer idealen Prognose auch als solche erkannt beziehungsweise vorhergesagt werden sollen.

Als Sensitivität lassen sich für das Zahlenbeispiel in der Vier-Felder-Tafel 80 Prozent und als Spezifität 90 Prozent errechnen. Demnach werden 80 Prozent der real Erkrankten durch den Test erkannt. Zugleich werden 90 Prozent der Gesunden auch korrekt als solche klassifiziert. Wie dies Ergebnis inhaltlich zu bewerten ist, hängt sehr maßgeblich vom Kontext ab, in dem der Test oder die Prognose eingesetzt werden soll. In Extremfällen könnte eine inhaltliche Bewertung dieses Ergebnisses im Hinblick auf die Praxistauglichkeit von „unbrauchbar“ bis „sehr hilfreich“ reichen.

Statistisch betrachtet unterscheiden sich die beiden genannten Werte zur Sensitivität und Spezifität zweifellos von Ergebnissen, die bei einer rein zufälligen Zuordnung von positiven und negativen Test- oder Prognose-Ergebnissen in der Population zu erwarten wären.[9] Im Hinblick auf Prognosen vieler gesundheitlicher Risiken dürfte die im Beispiel erreichte Kombination von Sensitivität und Spezifität als Hinweis auf eine vergleichsweise gute Prognose gelten können.

9 Hierbei wäre sowohl bei Test-negativen als auch bei Test-positiven Personen anteilig – und abgesehen von zufälligen Abweichungen – mit circa 10 Prozent wahr-Betroffenen in beiden Gruppen zu rechnen. Bei unveränderter Gesamtzahl an positiven und negativen Test-Ergebnissen wäre eine Besetzung der vier Zellen von A bis D mit 17, 153, 83 und 747 Personen zu erwarten, die dann auch den sogenannten Erwartungswerten des Chi2-Tests entsprechen. Bei dieser Besetzung ergibt sich eine Sensitivität von 17 und eine Spezifität von 83 Prozent. Der Chi2-Wert für die in der Abbildung dargestellte und von Erwartungswerten deutlich abweichende Zahlenkonstellation beträgt 312,5 und deutet auf hochsignifikante Abweichungen von einer zufällig erwartbaren Verteilung hin.

Eine wesentliche Eigenschaft der Sensitivität und Spezifität als Kennzahlen besteht darin, dass ihre Berechnung unabhängig von den realen Betroffenenanteilen in der jeweils konkret betrachteten Untersuchungspopulation ist. Unabhängig davon, ob in einer Population 1, 10 oder 50 Prozent aller Personen real betroffen sind, sollten Analysen in allen Populationen zumindest grundsätzlich dieselben Ergebnisse zur Sensitivität und Spezifität eines bestimmten Tests oder eines Prognosemodells liefern. Unterschiedliche Prävalenzen verändern lediglich die Zahlenverhältnisse zwischen den Spalten der Vier-Felder-Tafel in Tabelle 3-1, jedoch nicht die Zahlenverhältnisse innerhalb der beiden Spalten, auf die es bei der Berechnung der Sensitivität und Spezifität jeweils ausschließlich ankommt. **Sensitivität und Spezifität beschreiben Eigenschaften eines Tests oder einer Prognose (mit einem bestimmten Schwellenwert), die in Populationen mit unterschiedlichen Prävalenzen gleichermaßen Gültigkeit besitzen.** So sollten beispielsweise Sensitivitäten und Spezifitäten eines Tests auf das Vorliegen einer SARS-CoV-2-Infektion, die zu Beginn einer Pandemie mit weniger als einem Prozent Betroffenen in der untersuchten Population ermittelt wurden, grundsätzlich auch dann noch gültig sein, wenn in einer getesteten Population die Hälfte aller Personen real infiziert sind.

Demgegenüber bewerten die Kennzahlen PPV und NPV Eigenschaften von Tests oder Prognosen, die gemäß den nachfolgenden Definitionen von der Prävalenz des Ereignisses abhängen:

- Als **positiver Vorhersagewert (englisch: precision, auch positive predictive value – PPV)** eines Tests bezeichnet man den Anteil der real betroffen Personen unter allen positiv getesteten Personen. Der PPV gibt Auskunft darüber, wie viele der als positiv getesteten Personen anteilig auch real betroffen beziehungsweise erkrankt sind.
- Als **negativer Vorhersagewert (englisch: negative predictive value – NPV)** eines Tests bezeichnet man den Anteil der real nicht betroffenen Personen unter allen als negativ getesteten Personen. Der NPV gibt Auskunft darüber, wie viele der negativ getesteten Personen anteilig auch real nicht betroffen beziehungsweise gesund sind.

Insbesondere zur Beurteilung einer Eignung von Prognosemodellen oder von Tests für den praktischen Einsatz in bestimmten Populationen können der PPV sowie der NPV als Kennzahlen relevante Anhaltspunkte liefern. Beispielsweise lässt sich für das Zahlenbeispiel in Abbildung 3-1 ein PPV von 47 Prozent errechnen. Demnach muss bei nur weniger als die Hälfte der als positiv getesteten Personen in der betrachteten Beispielpopulation davon ausgegangen werden, dass sie auch real erkrankt sind, was für Betroffene und Behandler eine sehr relevante Information sein kann. Würde derselbe Test in einer Population eingesetzt, in der die Prävalenz bei nur ein Prozent liegt, würde ein PPV von nur noch 7,5 Prozent resultieren – nur noch weniger als ein Zehntel der Test-positiven wäre auch real erkrankt.[10] Derartige Ergebnisse können sehr wesentlich für die Entscheidung über den praktischen Einsatz eines Tests oder eine Prognosemodells in einer bestimmten Population sein. **Zur Abschätzung der allgemeinen Güte eines Tests oder einer Prognose ist der PPV aufgrund seiner Abhängigkeit von der Prävalenz jedoch nur bedingt geeignet.** Gleiches gilt für den NPV. Zudem lassen sich PPV und NPV bei bekannter Sensitivität und Spezifität auch unter Annahme einer bestimmten Prävalenz ohne eine erneute empirische Untersuchung für unterschiedliche Populationen abschätzen. Insofern erscheinen die beiden Kennwerte Sensitivität und Spezifität zur allgemeinen Charakterisierung der Güte eines medizinischen Tests oder eines Prognosemodells (nach Festlegung eines bestimmten Schwellenwertes) besser geeignet.

10 Das Ergebnis lässt sich leicht nachvollziehen, wenn man die Zahl der nicht Betroffenen in der rechten Spalte der Vier-Felder-Tafel in Abbildung 3-1. Vier-Felder-Tafel (engl. Confusion Matrix) zur Beurteilung der Güte von Vorhersagen (oben) und assoziierte Kennwerte (unten)

jeweils mit 11 multipliziert.

Tabelle 3-1. Kennwerte zu Güte von Vorhersagen und Tests – synonym verwendete Begriffe

Kennwert	Berechnung	alternative Bezeichnungen	englischsprachige Bezeichnungen
Sensitivität	TP / (TP + FN)	Richtig-Positiv-Rate	sensitivity, recall, true positive rate, hit rate, power
1 – Sensitivität	FN / (TP + FN)	Falsch-Negativ-Rate	false negative rate, miss rate
Spezifität	TN / (TN + FP)	Richtig-Negativ-Rate	specificity, true negative rate, correct rejection rate
1 – Spezifität	FP / (TN + FP)	Falsch-Positiv-Rate, Ausfallrate	false positive rate, fall-out
positiver Vorhersagewert	TP / (TP + FP)	positiver prädiktiver Wert, Genauigkeit, Relevanz, Wirksamkeit	positive predictive value (PPV), precision
negativer Vorhersagewert	TN / (TN + FN)	negativer prädiktiver Wert, Segreganz, Trennfähigkeit	negative predictive value (NPV)

TP: true positive, FP: false positive, FN: false negative, TN: true negative – vgl. auch Abbildung 3-1

3.2 F1-Score

Der F1-Score ist eine weit verbreitete Metrik im Bereich des maschinellen Lernens zur Bewertung der Leistung von binären Klassifikationsmodellen. Er ist das harmonische Mittel von Precision (positiver prädiktiver Wert, PPV) und Recall (Sensitivität) und bietet eine einzelne Metrik, die sowohl die Fähigkeit eines Modells, positive Instanzen korrekt zu identifizieren (Precision/PPV), als auch die Fähigkeit, alle relevanten positiven Instanzen zu finden (Recall/Sensitivität), ausbalanciert. Dies macht den F1-Score besonders nützlich in Szenarien mit unausgewogenen Datensätzen oder wenn sowohl falsche positive als auch falsche negative Ereignisse gleichermaßen wichtig sind.

Der F1-Score wird wie folgt berechnet:

$$F_1 = 2\frac{Precision \times Recall}{Precision + Recall}$$

Mit Precision = positiver prädiktiver Wert (PPV), Recall = Sensitivität – vgl. Abbildung 3-1

Der F1-Score reicht von 0 bis 1, wobei 1 perfekte Precision und Recall zeigt, während 0 bedeutet, dass entweder Precision oder Recall null ist. Ein hoher F1-Score zeigt an, dass das Modell ein gutes Gleichgewicht zwischen Precision und Recall erreicht hat, was ihn zu einer wertvollen Metrik für die Bewertung der Leistung von Klassifikationsmodellen macht.

Zwei Hauptmerkmale unterscheiden den F1-Score vom MCC. Der F1-Score variiert, wenn die Merkmalsausprägungen vertauscht, d. h. die Klasse der Outcomes invers codiert, werden, während der MCC dabei invariant ist. Dieses Problem kann durch die Anwendung des Mikro/Makro-Durchschnittsverfahrens auch im binären Fall überwunden werden, indem der F1-Score sowohl für die positive als auch für die negative Klasse definiert und dann die beiden Werte gemittelt werden (Makro), sowie unter Verwendung des durchschnittlichen Recalls und der durchschnittlichen Precision (Mikro). Der mikro/makro-gemittelte F1 ist invariant gegenüber Klassenvertauschung und sein Verhalten ähnelt mehr dem MCC. Zudem ist der F1-Score unabhängig von der Anzahl der korrekt als negativ klassifizierten Proben. Trotz einiger Kritik bleibt der F1-Score eine der am weitesten verbreiteten Metriken unter Forschern im Bereich des maschinellen Lernens (Chicco & Jurman, 2020).

Betrachten wir als Beispiel ein positiv unausgeglichenes Datenset, das aus 91 gesunden Personen und 9 kranken Patienten (Prävalenz = 9 %) besteht. Angenommen, das Prognosemodell erzeugt die folgenden Werte: TP = 90, FN = 1, TN = 0, FP = 9. In diesem Fall besticht das Prognosemodell mit seiner Fähigkeit, die positiven Dateninstanzen vorherzusagen (90 gesunde Patienten von 91 wurden korrekt vorhergesagt). Gleichzeitig zeigt es aber auch eine eingeschränkte Aussagekraft bei der Identifizierung kranker Personen (keine Person wurde korrekt vorhergesagt). Trotzdem zeigen Genauigkeit (Accuracy) und F1-Wert in diesem Fall hohe Werte: Genauigkeit = 0,90 und F1-Score = 0,95, beide nahe dem bestmöglichen Wert von 1,00 im Intervall [0, 1]. Der MCC hingegen liegt bei MCC = -0,03.

3.3 Matthews-Korrelationskoeffizient

Der Matthews-Korrelationskoeffizient (MCC) ist eine Metrik zur Bewertung der Leistung von Algorithmen, die binäre Ergebnisse vorhersagen (Matthews, 1975). Er basiert auf der Vier-Felder-Tafel (engl. Confusion Matrix), in der die Anzahl richtig und falsch positiver, sowie richtig und falsch negativer Ergebnisse eines Prognosemodells verzeichnet sind (s. Abbildung 3-1) und wird mit folgender Formel berechnet:

$$MCC = \frac{TN \times TP - FN \times FP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Mit TP: true positive, FP: false positive, FN: false negative, TN: true negative – vgl. auch Abbildung 3-1

Die Werte des MCC reichen von -1 bis 1. Ein Wert von 1 zeigt eine perfekte Vorhersage, 0 eine zufällige Vorhersage und -1 eine komplette Diskrepanz zwischen Vorhersage und tatsächlichem Ergebnis. Der MCC gilt als zuverlässige Metrik, da er alle vier Zellen der Vier-Felder-Tafel berücksichtigt und somit einen umfassenden Überblick über die Modellleistung bietet. Dies ist besonders wertvoll in Szenarien mit unausgewogenen Klassen, da er Einblicke in die Fähigkeit des Modells bietet, sowohl positive als auch negative Instanzen korrekt zu identifizieren. Im Vergleich zu anderen Metriken wie Genauigkeit, PPV (Precision), Sensitivität (Recall) und dem F1-Score bietet der MCC ein differenzierteres Verständnis der Modellleistung, besonders bei unausgewogenen Datensätzen.

Trotz seiner Vorteile gibt es jedoch Situationen, in denen der MCC nicht definiert werden kann oder große Schwankungen aufweist, besonders bei extrem unausgeglichenen Klassifikationen (Chicco & Jurman, 2020).

3.4 AUC-ROC – Fläche unter der Receiver Operating Characteristic

In der Regel liefern Prädiktionsmodelle primär keine Vorhersagen im Sinne der beiden dichotomen und komplementären Kategorien „ist betroffen“ oder „ist nicht betroffen“. Stattdessen werden in den Modellen für Personengruppen mit bestimmten Merkmalskonstellationen primär zumeist vorhergesagte Wahrscheinlichkeiten für den Eintritt von zuvor definierten Ereignissen ausgegeben (z. B. die Eintrittswahrscheinlichkeit von X Prozent für ein vorhergesagtes Ereignis in einer bestimmten Personengruppe). Um für Prädiktionsmodelle mit derartigen Vorhersagewerten Sensitivitäten und Spezifitäten berechnen zu können, ist grundsätzlich die Festlegung eines Schwellenwertes (threshold) erforderlich, der Personen aufgrund der im Modell vorhergesagten Wahrscheinlichkeit jeweils der einen oder anderen Kategorie zuordnet.

Für derartige Kategorisierungen existieren jedoch keine allgemeingültigen Regeln, die stets zu auch inhaltlich optimalen Festlegungen von Grenzwerten führen. Die zunächst naheliegende Wahl eines Wertes von 0,5 beziehungsweise 50 Prozent als Schwellenwert ist in vielen Fällen nicht sinnvoll, beispielsweise da bei eher seltenen medizinischen Risiken in vielen Vorhersagemodellen für kaum eine oder auch keine der Konstellationen von Merkmalsausprägungen ein derart hohes Risiko ermittelt werden kann. Insofern bleiben die Festlegung eines entsprechenden Grenzwertes und damit auch die ermittelten Werte zur Sensitivität und Spezifität zwangsläufig abhängig von willkürlichen Entscheidungen.

Eine Möglichkeit, die Güte einer Vorhersage unabhängig von der Wahl eines Schwellenwertes zur Bestimmung der Sensitivität und Spezifität bestimmen zu können, bietet die Berechnung einer Kennzahl, in der alle sinnhaft möglichen Trennungen einer Population zur Bestimmung von Sensitivität und Spezifität zugleich berücksichtigt werden (Hajian-Tilaki, 2013). Zur Berechnung dieser Kennzahl wird zunächst eine sogenannte Receiver Operating Characteristic (ROC-Kurve, auch Grenzwertoptimierungskurve) ermittelt und als Kennzahl anschließend die Fläche unter dieser Kurve (engl.: „area under the curve“ – AUC) bestimmt.

Zur Ermittlung der ROC-Kurve werden zunächst alle Beobachtungen (Personen) absteigend nach den vorhergesagten Ereigniswahrscheinlichkeiten sortiert. Anschließend werden alle Kombinationen aus Richtig-Positiv-Raten (Sensitivitäten) und Falsch-Positiv-Raten (den komplementären Anteilen der Spezifitäten) ermittelt, die sich bei einer Berücksichtigung zunehmender Anteile der absteigend nach Ereigniswahrscheinlichkeiten sortierten Gruppen von Beobachtungen ergeben. Richtig-Positiv-Raten (Sensitivitäten) werden auf der y-Achse über die mit zunehmender Berücksichtigung von Beobachtungen zugleich größer werdenden Falsch-Positiv-Raten auf der x-Achse in einer Grafik als ROC-Kurve aufgetragen (vgl. Abbildung 3-2).

Da zu Beginn des Vorgangs keine Beobachtung als richtig positiv oder falsch positiv klassifiziert ist, beginnt die ROC-Kurve stets und ohne Ausnahme im Ursprung des Koordinatensystems (mit x = 0 und y = 0). Sinngemäß entspricht dies der Wahl eines Schwellenwertes, bei dem alle Beobachtungen als gesund klassifiziert werden: Niemand ist dann fälschlich positiv klassifiziert (womit die Spezifität bei 100 Prozent und die Falsch-Positiv-Rate bei 0 Prozent liegt), allerdings ist auch Niemand richtig positiv klassifiziert (womit auch die Sensitivität bei 0 Prozent liegt). Das umgekehrte Extrem markiert das Ende der ROC-Kurve: Werden durch die Wahl des Schwellenwertes schließlich alle Beobachtungen als krank klassifiziert, beträgt die Falsch-Positiv-Rate 1 beziehungsweise 100 Prozent (da auch alle Gesunden als krank klassifiziert wurden, womit die Spezifität dann den Wert 0 erreicht). Zugleich werden zwangsläufig alle real kranken Personen auch als krank klassifiziert, womit die Sensitivität beziehungsweise Richtig-Positiv-Rate stets 1 beziehungsweise 100 Prozent erreicht. Jede ROC-Kurve endet insofern im Punkt mit x = 1 und y = 1, also bildlich gesprochen rechts oben im Koordinatensystem. **Die ROC-Kurve**

stellt dar, wie sich die Sensitivität unter konsekutiver Verwendung aller basierend auf Ergebnissen einer Modellrechnung nur möglichen Schwellenwerten verändert – abgelesen werden kann aus der Kurve insofern, welche Falsch-Positiv-Rate in Kauf genommen werden muss, um eine bestimmte Sensitivität zu erreichen.

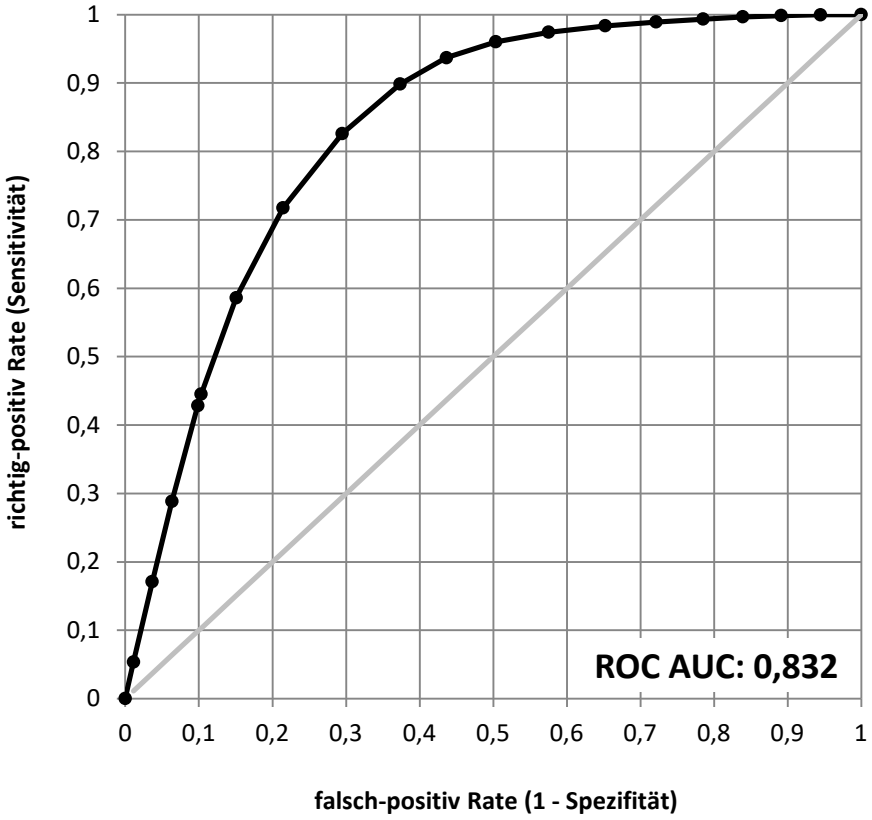

Abbildung 3-2. ROC-Kurve. Beispiel: Vorhersage des Diabetes-Risikos bei Männern abhängig vom Alter nach Gruppierung in 19 Altersgruppen; Diagnoseprävalenz altersübergreifend: 10,5 %

Abbildung 3-2 zeigt die ROC-Kurve zu einem einfachen, realitätsnahen Datenbeispiel. Dargestellt ist, wie sich das Risiko für das Vorliegen einer Diabetes-Diagnose bei Männern basierend auf Angaben ihrer Zuordnung zu 19 Altersgruppen vorhersagen lässt (Altersgruppen 0, 1 bis 4, 5 bis 9, 10 bis 14, …, 85 bis 89 sowie 90 und mehr Jahre). Da lediglich ein Merkmal im Sinne eines Prädiktors mit 19 Merkmalsausprägungen existiert, können maximal 19 Populationssubgruppen mit unterscheidbaren vorhergesagten Risiken und entsprechend viele Punkte auf der ROC-Kurve (zuzüglich des Punktes im Ursprung des Koordinatensystems) existieren. Der erste Datenpunkt bei x = 0,011 und y = 0,054 resultiert hier aus den Daten zu derjenigen Altersgruppe mit dem höchsten Diabetes-Risiko (im Beispiel sind dies 85- bis 89-jährige Männer mit einer Diabetes-Prävalenz von 36,1 Prozent). Gemäß dem Ergebnis zur Sensitivität umfasst diese Gruppe in der betrachteten Population 5,4 Prozent aller Diabetiker und gemäß der Falsch-Positiv-Rate 1,1 Prozent aller Nicht-Diabetiker. Die folgenden Punkte der Kurve resultieren dann durch konsekutive Einbeziehung von Personen aus den verbleibenden Subgruppen in einer absteigenden Sortierung nach Risiken, womit sowohl die Sensitivität als auch die Falsch-Positiv-Rate kontinuierlich steigen.

Würden Personen in zufälliger Reihenfolge, also ohne Prädiktion des Risikos, berücksichtigt, wäre – abgesehen von zufallsbedingten Effekten – eine übereinstimmende, d. h. proportionale Zunahme der Sensitivität und Falsch-Positiv-Rate zu erwarten, also eine ROC-Kurve, die der in Abbildung 3-2 eingezeichneten Diagonalen entspricht. Die Fläche unter dieser Diagonalen beträgt 0,5. Ein AUC-ROC-Wert

von 0,5 kennzeichnet insofern ein Ergebnis, das ohne eine effektive Risikovorhersage im Sinne eines sogenannten Nullmodells rein zufällig zu erwarten ist.

Die im Beispiel ermittelte Fläche unter der Kurve mit ROC AUC = 0,832 deutet demgegenüber auf eine Prädiktion hin, die deutlich von rein zufällig erwartbaren Ergebnissen abweicht und nach gängigen Kriterien bzw. Einschätzungen im medizinischen Bereich auf eine gute Prädiktionsgüte hindeutet. Im Falle einer perfekten Vorhersage würde ein ROC AUC = 1 resultieren – alle real Betroffenen würden identifiziert, ehe eine Person fälschlich als positiv klassifiziert wird, womit die Sensitivität bereits zu Beginn der Kurve den Wert 1 erreichen würde. Ein ROC AUC-Wert deutlich unterhalb von 0,5 oder nahe 0 als anderes Extrem deutet demgegenüber auf falsch kodierte Merkmalsausprägungen hin.

Tabelle 3-2 gibt einen Überblick zu üblichen Bewertungen von ROC AUC-Werten. Wie bei allen allgemeinen Kategorisierungen von Kennwerten ist auch hier zu beachten, dass inhaltliche Bewertungen immer auch den spezifischen Kontext berücksichtigen sollten, in dem Ergebnisse ermittelt wurden und Bewertungen dann im Einzelfall merklich abweichen können.

Tabelle 3-2. Bewertung von ROC AUC-Werten nach Hosmer, Lemeshow, and Sturdivant (2013)

ROC AUC-Wertebereich	Bewertung
1,0	perfekte Prädiktion
über 0,9	Hervorragend
über 0,8	Ausgezeichnet
über 0,7	Akzeptabel
um 0,5	keine Prädiktion, zufällig angeordnete Beobachtungen
deutlich unter 0,5	Hinweis auf falsch kodierte Merkmalsausprägungen

3.5 AUC-PR – Fläche unter der Precision-Recall-Curve

Neben der im Bereich der Epidemiologie zur Beurteilung der Prädiktionsgüte von Modellen seit Jahrzehnten etablierten ROC-Kurve wird insbesondere im Bereich des maschinellen Lernens häufiger auch eine andere Kurve betrachtet – die sogenannte Precision-Recall-Curve (PR-Kurve), wobei auch hier Wertepaare genutzt werden, die zuvor aus Ergebnissen zu einer absteigend nach vorhergesagten Ereigniswahrscheinlichkeiten sortierten Population ermittelt wurden. Bei dieser Kurve wird der positive Vorhersagewert (PPV, precision) auf der y-Achse über der Richtig-Positiv-Rate (Sensitivität, recall) auf der x-Achse aufgetragen (vgl. Abbildung 3-3).

Es wird demnach dargestellt, wie sich der der PPV (im Sinne des Anteils der real Betroffenen unter den Test-Positiven) bei Verschiebung des Schwellenwertes mit dann zunehmender Sensitivität verhält. Bei geringer Sensitivität (zu Beginn der Kurve) und einer Berücksichtigung ausschließlich derjenigen Personen mit den höchsten vorhergesagten Risiken sollte dabei der PPV die höchsten Werte aufweisen – in der Regel sollte die Kurve also tendenziell von „links oben“ beginnen. Zumindest bei Daten mit sehr vielen Beobachtungen und nicht gänzlich perfekten Vorhersagen ist zudem zu erwarten, dass eine Sensitivität von 100 Prozent erst nach Berücksichtigung weitgehend aller Beobachtungen erreicht wird. Ist dies der Fall, entspricht der PPV am Ende der Kurve dann erwartungsgemäß der anteiligen Betroffenenrate in der Gesamtpopulation oder liegt geringfügig darüber. Sinngemäß entspricht dies dem PPV einer Vorhersage, die die gesamte betrachtete Population „wahllos“ für betroffen erklärt. Dann sind mit dieser Aussage anteilig genauso viele Personen real betroffen, wie dies bereits die Betroffenenrate zur Gesamtpopulation zum Ausdruck bringt.

Dies bedeutet, dass die PR-Kurve bei einer Ereigniswahrscheinlichkeit von 0,5 in der betrachteten Gesamtpopulation auf der rechten Seite erwartungsgemäß in Höhe von etwa 0,5 endet. Werden demgegenüber Vorhersagen von Ereignissen betrachtet, von denen ein Prozent der Gesamtpopulation betroffen ist, endet die Kurve erwartungsgemäß in Höhe von 0,01 – die Höhe des Endes der PR-Kurve wird ausschließlich durch die Ereigniswahrscheinlichkeit in der betrachteten Gesamtpopulation bestimmt und sagt damit nichts über die Güte der Prädiktion aus.

Entsprechend variiert – unabhängig von der Prädiktionsgüte eines Modells – auch die ermittelte Fläche unter der PR-Kurve (AUC-PR). Bei einer Betroffenenrate von 90 Prozent gilt auch bei wahlloser Deklaration der Gesamtpopulation als betroffen noch ein PPV von 0,9 mit einer entsprechenden Fläche unter der Kurve erreicht, während sich bei einer Betroffenenrate von 1 Prozent dann lediglich ein PPV sowie eine Fläche von 0,01 berechnen lässt. Vergleiche von AUC-PR bzw. den darunter liegenden Flächen können nur dann direkt zum Vergleich der Güte von zwei Prädiktionsmodellen herangezogen werden, wenn sich diese auf Vorhersagen in derselben Population oder in zwei Populationen mit übereinstimmenden Ereigniswahrscheinlichkeiten beziehen.

Der Wert der AUC-PR liegt insofern eher darin, den praktischen Nutzen eines Prädiktionsmodells in einer Population mit einer bestimmten erwarteten Ereigniswahrscheinlichkeit beurteilen zu können. So lässt sich beispielsweise aus der PR-Kurve in Abbildung 3-3 zum bereits erläuterten Beispiel zur Vorhersage eines Diabetes-Diagnose-Risikos ablesen, dass bei den dargestellten Prädiktionsergebnissen für die Detektion von mehr als der Hälfte aller real Betroffenen die Wahl eines Schwellenwertes erforderlich ist, bei dem dann noch ein PPV von knapp über 0,3 erreicht wird. In einer derart selektierten Risikogruppe wären also nur knapp über 30 Prozent auch real von einem Diabetes betroffen. Dies könnte dann möglicherweise in der Praxis noch ein ausreichend hohes Risiko sein, um die Verteilung von Aufklärungsbroschüren zum Thema Diabetes zu rechtfertigen, wäre aber vermutlich keine Begründung dafür, an diese Gruppe ohne eine weitere Eingrenzung kostspielige Geräte zur regelmäßigen Bestimmung des Blutzuckerspiegels zu verteilen.

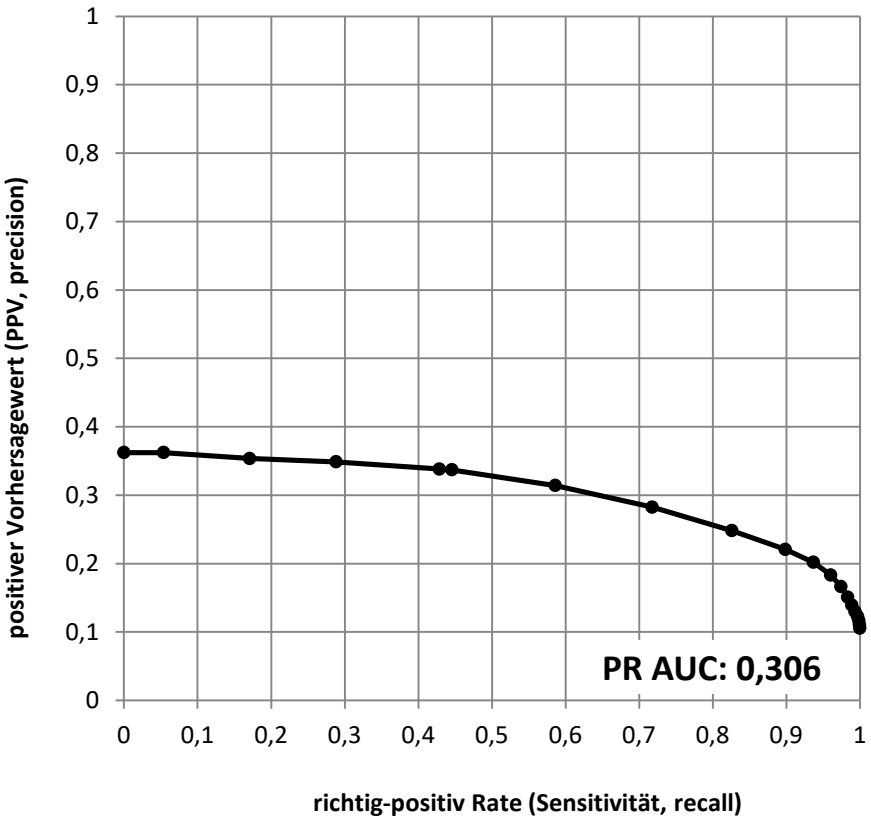

Abbildung 3-3. Precision-Recall-Curve. Beispiel: Vorhersage des Diabetes-Risikos bei Männern abhängig vom Alter nach Gruppierung in 19 Altersgruppen; Diagnoseprävalenz altersübergreifend: 10,5 %

Precision Recall-Kurven – Darstellungsvarianten und Flächenberechnung

Während sich Darstellungen von ROC-Kurven grundsätzlich fast immer ähneln, finden sich im Hinblick auf PR-Kurven bei Recherchen im Internet sehr unterschiedlich anmutende Darstellungsvarianten. Häufiger als die für Abbildung 3-3 gewählte Form mit einer bei steigendem Recall stetig abnehmenden Precision lassen sich Kurven mit einem eher „sägezahnartigen“ Verlauf identifizieren. **Offensichtlich wird das Vorgehen bei der Darstellung von PR-Kurven in der Praxis unterschiedlich gehandhabt, was die Interpretation vorgefundener Kurven erheblich erschwert.** Ähnliches könnte auch für die Berechnung der AUC-PR, also der Fläche unter der Precision Recall-Kurve, in manchen Programmen gelten. Im Anhang 11 wird näher auf Varianten eingegangen, die die bei der Darstellung von PR-Kurven und der zugehörigen AUC-Berechnung existieren.

Quellen

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*(1), 6. doi:10.1186/s12864-019-6413-7

Hajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian J Intern Med, 4*(2), 627-635.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. New York: John Wiley & Sons.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure, 405*(2), 442-451.

4 Aufbereitung von Routinedaten für konventionelle und ML-basierte Vorhersagen

Bevor die eigentliche Analyse von Routinedaten durchgeführt werden kann, müssen die Daten i.d.R. in mehreren Schritten aufbereitet werden.
Typischerweise erfolgt nach der Datenlieferung zunächst eine Prüfung, ob die Daten entsprechend der Spezifikation geliefert wurden. Weitere Aufbereitungsschritte zielen etwa auf die Reduktion des Speicherbedarfs, die Aufbereitung von zeitlich überlappenden und redundanten Intervallen, den Umgang mit zensierten oder fehlenden Daten und die Erstellung des Analysedatensatzes für die eigentlichen Analysen ab.

Durch die gesetzlichen Regelungen zum elektronischen Datenaustausch in der GKV (s. Kapitel 1) liegen die Routinedaten bei allen gesetzlichen Krankenkassen in einem strukturierten und relativ gut vergleichbaren Format vor. Dies ist über einheitlich definierte Merkmalsausprägungen, wie Schlüssel oder Klassifikationssysteme, gewährleistet (s. Kapitel 1.7). Unstrukturierte Daten, beispielsweise in Form von Freitextfeldern, bilden die Ausnahme. Die Aufbereitung der Daten für Prognosemodelle ist daher i.d.R. mit weniger Aufbereitungsschritte bzw. mit weniger umfangreichen Schritten verbunden im Vergleich zu unstrukturierten (unterannotierten) Daten, die z. B. in Form von Texten, Bildern oder Videos vorliegen. Solche unstrukturierten Daten fallen an vielen Stellen im Gesundheitssystem an, wie etwa bei Arztbriefen, Notizen von Ärzten/Ärztinnen aus Patientengesprächen, Daten aus bildgebenden Verfahren wie Röntgen oder CT oder Daten aus Sensoren oder anderen medizinischen Messgeräten. Trotzdem sind auch bei den Routinedaten der Krankenkassen Aufbereitungsschritte notwendig, bevor die Daten wie geplant analysiert werden können. Art und Umfang der Aufbereitungsschritte sind dabei abhängig vom jeweiligen Auswertungsvorhaben. Trotzdem sollen in diesem Kapitel typische Aufbereitungsschritte benannt und beispielhaft erläutert werden. Als Praxisbeispiel werden dazu die Daten aus dem Projekt KI-THRUST verwendet, die in den nachfolgenden Kapiteln zum Vergleich von konventionellen und KI-basierten Methoden herangezogen werden. Mit dieser Orientierung an einem realen Datenbeispiel sollen insbesondere auch Anhaltspunkte vermittelt werden, in welchem Umfang bestimmte Daten bereitstehen. Daher folgt zunächst eine kurze Einführung in das KI-THRUST Projekt.

4.1 Einführung in das Praxisbeispiel KI-THRUST

Das Ziel des Projekts KI-THRUST ist ein Vergleich von konventionellen und KI-basierten Vorhersagemodellen an einem praxisorientierten Beispiel, nämlich der Vorhersage von Nachsorgebedarfen und unerwünschten Ereignissen im Anschluss an einen Krankenhausaufenthalt. Im Folgenden wird das Projekt kurz vorgestellt. Im weiteren Verlauf des Weißbuches wird das Beispiel KI-THRUST referenziert, um Aufbereitungs- und Analyseschritte zu veranschaulichen und die Ergebnisse des Vergleichs zwischen konventionellen und KI-basierten Vorhersagemodellen vorzustellen.

Im Projekt lagen Daten von Versicherten der Betriebskrankenkassen (BKK) vor, die in den Jahren 2015 bis 2020 mindestens einmal aus einem stationären Krankenhausaufenthalt entlassen wurden. Die Daten umfassen Informationen von knapp 1,4 Millionen Versicherten zu Stammdaten, Versicherungszeiten, ambulanten Behandlungen, stationären Behandlungen, Heilmitteln, Arzneimitteln, Hilfsmittel, Rehabilitation und dem Pflegegrad. Die Daten lagen pseudonymisiert in verschiedenen Tabellen vor (s. Tabelle 4-1). Die Verknüpfung zwischen den Tabellen wurde über ein Versichertenpseudonym ermöglicht, sowie innerhalb eines Behandlungsfalls im ambulanten und stationären Bereich zusätzlich über eine jeweils vorhandene Fallnummer.

Tabelle 4-1 Informationen zu verfügbaren Datentabellen im Projekt KI-THRUST. Enthalten sind Daten von allen BKK-Versicherten aus den Jahren 2015 bis 2020, die in diesem Zeitraum mindestens eine Entlassung aus einem stationären Krankenhausaufenthalt hatten.

Nr.	Beschreibung	Enthaltene Variablen	Anzahl Einträge in Tabelle
1	Stammdaten der Versicherten	Versicherten-ID Geburtsjahr Todesdatum Geschlecht Postleitzahl	1.423.310 Datenzeilen 1.423.310 Versicherte 1 Eintrag pro Versicherten
2	Versicherungszeiten	Versicherten-ID Beginn Ende	1.681.039 Datenzeilen 1.423.310 Versicherte 1-303 Einträge pro Versicherten
3	Ambulante Behandlungen: Falltabelle	Versicherten-ID Fallnummer Berichtsjahr Berichtsquartal BSNR Beginn Fall Ende Fall Art der Inanspruchnahme	86.906.613 Datenzeilen 86.906.613 ambulante Fälle 1.419.690 betroffene Versicherte ~ 55 Fälle pro Betroffenen (Range: 1-984)
4	Ambulante Behandlungen: Diagnosen	Versicherten-ID Fallnummer ICD Diagnosesicherheit Seitenlokalisation Diagnose	440.892.079 Datenzeilen 71.823.423 Fälle (etwas geringer als in der ambulanten Falltabelle, da einzelne Fälle ohne Diagnoseangabe auftauchen, z. B. bei Laborärzten/-ärztinnen) ~ 4 Diagnosen pro Fall (Range: 1-3.976) 1.419.593 Versicherte

Nr.	Beschreibung	Enthaltene Variablen	Anzahl Einträge in Tabelle
5	Ambulante Behandlungen: EBM-Ziffern	Versicherten-ID Fallnummer LANR Gebührenordnungsnummer Datum der Leistung	702.722.409 Datenzeilen 86.642.283 Fälle (sollte nahezu identisch mit der Fallzahl in der ambulanten Fall-tabelle sein bis auf einzelne Fehler und Ausnahmefälle) ~ 6 Einträge pro Fall (Range: 1-684) 1.419.689 Versicherte
6	Ambulante Behandlungen: OPS-Daten	Versicherten-ID Fallnummer OPS-Code Seitenlokalisation OPS	2.165.564 Datenzeilen 1.361.302 Fälle ~ 1 Eintrag pro Fall (Range: 1-82) 525.780 Versicherte
7	Arzneimittelverordnungen	Versicherten-ID BSNR verordnender Arzt LANR verordnender Arzt PZN ATC-Code DDD Datum der Verordnung	108.540.928 Datenzeilen 1.406.013 Versicherte ~ 48 Einträge pro Versicherten (Range: 1-3.345)
8	Stationäre Behandlungen: Falltabelle	Versicherten-ID Fallnummer IK Aufnahmedatum Entlassdatum Aufnahmegrund Entlassgrund DRG	4.090.658 Datenzeilen 4.090.658 Fälle 1.423.310 Versicherte ~ 2 Einträge pro Versicherten (Range: 2-258)
9	Stationäre Behandlungen: Diagnosen	Versicherten-ID Fallnummer Diagnose Seitenlokalisation Diagnose Diagnoseart Sekundäre Diagnose Seitenlokalisation sek. Diag-nose Fachabteilung	36.106.938 Datenzeilen 4.090.657 Fälle (sollte nahezu identisch mit der Fallzahl in der stationären Fallta-belle sein bis auf einzelne Fehler) ~ 7 Einträge pro Fall (Range: 1-97) 1.423.310 Versicherte
10	Stationäre Behandlungen: OPS	Versicherten-ID Fallnummer Operationstag OPS-Code Seitenlokalisation OPS	17.808.154 Datenzeilen 3.588.013 Fälle ~ 3 Einträge pro Fall (Range: 1-501) 1.315.810 Versicherte
11	Heilmittel	Versicherten-ID Fallnummer BSNR verordnender Arzt LANR verordnender Arzt	56.658.905 Datenzeilen 755.136 Versicherte ~ 29 Einträge pro Betroffenen (Range: 1-9.162)

Nr.	Beschreibung	Enthaltene Variablen	Anzahl Einträge in Tabelle
		Fachgruppe verordnender Arzt Verordnungsdatum Indikationsschlüssel Heilmittelpositionsnummer Datum Leistungserbringung	
12	Hilfsmittel	Versicherten-ID BSNR verordnender Arzt LANR verordnender Arzt Fachgruppe verordnender Arzt Verordnungsdatum Hilfsmittelpositionsnummer Datum Leistungserbringung Fachgruppe	12.264.226 Datenzeilen 927.712 Versicherte ~ 5 Einträge pro Betroffenen (Range: 1-2.193)
13	Pflege	Versicherten-ID Pflegegrad Pflegegrad ab Datum Pflegegrad bis Datum	322.099 Datenzeilen 204.408 Versicherte ~ 1 Eintrage pro Betroffenen (Range: 1-7)
14	Reha	Versicherten-ID Fallnummer BSNR / IK Beginn Ende Art (Freitext, je nach Krankenkasse, z. B. „ambulante Reha / Geriatrie“) Diagnose	123.861 Datenzeilen 107.236 Versicherte ~ 1 Eintrag pro Betroffenen (Range: 1-7)

In KI-THRUST wurde untersucht, wie gut der Nachsorgebedarf von Patienten mit bereits zu Beginn des Krankenhausaufenthalts verfügbaren Daten vorhergesagt werden kann.

Es wurden zwei Outcomes ausgewählt, die mithilfe der Modelle vorhergesagt werden sollten und die im Rahmen von vorangegangenen Projekten (EMSE und USER, gefördert durch den Innovationsausschuss beim Gemeinsamen Bundesausschuss unter den Kennzeichen 01VSF16041 und 01NVF18010) als relevant für die Krankenhausnachsorge identifiziert wurden:

- **Mortalität**: Versterben innerhalb von 30 Tagen nach Entlassung aus dem Krankenhaus,
- **Ungeplante Wiederaufnahmen**: Stationäre Wiederaufnahmen ins Krankenhaus innerhalb von 30 Tagen nach Entlassung mit dem Aufnahmegrund „Notfall“.

4.2 Prüfung der gelieferten Daten

Bevor die Daten aufbereitet werden, sollte zunächst überprüft werden, ob sie entsprechend der zuvor vereinbarten Spezifikation geliefert wurden. Eine gezielte Prüfung der Daten in verschiedenen Bereichen kann eventuelle Fehler in der Datenübertragung oder Abstimmung aufdecken.

Typischerweise werden Daten für Forschungszwecke von den eigentlichen Datenhaltern (z. B. Krankenkassen) selektiert und bereitgestellt. Bevor die Daten aufbereitet und Analysen durchgeführt werden können, sollte überprüft werden, ob die Daten entsprechend den ursprünglichen Anforderungen geliefert wurden. Auch bei größter Sorgfalt kann es immer dazu kommen, dass eine Datenlieferung fehlerhaft ist, weil beispielsweise die Datenextraktion oder die Datenübertragung unerwartet abbricht oder weil es Unklarheiten darüber gibt, wie die Daten selektiert werden sollen.

Eine umfassende Datenprüfung ist immer individuell und hängt vom jeweiligen Fall ab. Dennoch gibt es einige Aspekte, die typischerweise zu prüfen sind und die im Folgenden erläutert werden. Dabei werden zunächst grundlegende und allgemeine Datenprüfungen vorgestellt, um dann auf komplexere Prüfmöglichkeiten einzugehen, die abhängig von den Daten und deren zeitlicher Verfügbarkeit durchgeführt werden können.

Tabelle 4-2 Grundlegende Fragen bei der Prüfung bereitgestellter Routinedaten

Überprüfter Abschnitt	Beschreibung der Prüfung
Datentabellen	Sind alle angefragten Datentabellen vorhanden?
Variablen/Features	Sind alle angeforderten Variablen/Features in den jeweiligen Datentabellen enthalten und liegen sie im spezifizierten Format vor?
Klassifikations-/Schlüsselverzeichnisse	Sind alle relevanten Klassifikations- und Schlüsseltabellen verfügbar? Neben den häufig genutzten Klassifikationssystemen wie ICD oder EBM (s. Kapitel 1.7) können je nach Krankenkasse auch weitere Systeme genutzt werden, z. B. solche mit Behandlungsziffern aus Selektivverträgen. Um die Daten in diesen Fällen auswerten zu können, werden die entsprechenden Schlüsseldateien benötigt.
Datenintegrität	Ist die Datenintegrität gewahrt, d. h. sind die verschiedenen Tabellen – sofern vorgesehen – über Schlüsselvariablen, wie Versicherten-ID oder Fall-Nummer, verknüpfbar?
Datenmenge	Entspricht die Anzahl an Beobachtungen in den Datentabellen den Erwartungen, z. B. die Anzahl an Versicherten oder die Anzahl der Fälle? In der Regel kann keine genaue Anzahl vorhergesagt werden, aber aufgrund von Erfahrungswerten kann man abschätzen, wie hoch beispielsweise der Anteil an Versicherten mit stationärem Aufenthalt oder ambulanten Behandlungen ungefähr sein sollte (s. dazu auch die Übersichtskästen in Kapitel 1 zu den verschiedenen Datentabellen). Extreme Abweichungen von dieser Schätzung deuten auf eine unvollständige Datenlieferung hin.
Merkmalsausprägungen	Sind alle zu erwartenden Merkmalsausprägungen bei den Variablen in der spezifizierten Form vorhanden (z. B. Variable „Geschlecht“ mit

Überprüfter Abschnitt	Beschreibung der Prüfung
	den Ausprägungen: „M“/„W“/„D“)? Gibt es eine große Anzahl an unerwartet fehlenden Werten oder ungültige/unerwartete Merkmalsausprägungen? Es kann immer wieder vereinzelt zu fehlenden Werten kommen, insbesondere bei bestimmten Variablen. Auffällig ist aber, wenn sich fehlende Daten systematisch häufen, insbesondere für einzelne Subgruppen oder Cluster (wie Datenjahre, Altersgruppen, Einrichtungen oder räumliche Einheiten).
Erhebungszeiträume	Sind die Daten in unterschiedlichen Erhebungszeiträumen wie erwartet vorhanden? Gibt es z. B. eine vergleichbare Menge an Daten für alle spezifizierten Jahre, Monate oder Quartale?
Häufigkeiten von Beobachtungen und Häufigkeitsverteilung von Merkmalsausprägungen	Häufig bietet sich auch eine Plausibilitätskontrolle der Daten an, zumindest für die relevanten Merkmale. Dies ist abhängig von der jeweiligen Fragestellung. Beispielsweise könnte überprüft werden, ob die Alters- oder Geschlechterverteilung im Allgemeinen und bei bestimmten Diagnosen in etwa den Erwartungen bzw. anderen Quellen (Studien, amtlichen Statistiken, etc.) entspricht. Wenn Daten von verschiedenen Quellen geliefert werden, z. B. verschiedenen bereitstellenden Krankenkassen oder verschiedenen Kassenärztlichen Vereinigungen, können Fehler einzelner Lieferungen/Quellen in der Gesamtbetrachtung der Daten untergehen. Daher lohnt sich auch eine getrennte Betrachtung der oben genannten Aspekte für alle Einzellieferungen.

4.3 Aufbereitung der Daten für die Analysen

Einige Aufbereitungsschritte werden häufig bei Routinedaten durchgeführt, bevor der eigentliche Analysedatensatz erstellt werden kann. Zunächst kann man prüfen, ob sich eine Reduktion des Speicherbedarfs der Daten lohnt. Dann müssen u.U. zeitlich überlappende oder redundante Intervalle aufbereitet werden. Bei zensierten oder fehlenden Daten muss überprüft werden, ob Daten ausgeschlossen oder z. B. imputiert werden sollen.

Nach der initialen Prüfung der Daten folgt in der Regel die Aufbereitung der Daten für die nachfolgenden Analysen. Auch hier werden allgemeine Aufbereitungsschritte, die in vielen Anwendungsfällen relevant sind, kurz erläutert.

4.3.1 Speicherbedarf der Daten reduzieren

Routinedaten können je nach Größe der Versichertenpopulation und Umfang der benötigten Tabellen mehrere Gigabyte umfassen. Die Verarbeitung der Daten kann daher zeitlich sehr aufwendig werden und es lohnt sich i.d.R., die Daten zu verkleinern. Eine Möglichkeit, das Datenvolumen ohne Informationsverlust zu verkleinern, kann sein, längere alphanumerische Variablen (mit mehr als 8 Zeichen) möglichst durch numerische Variablen zu ersetzen. Ein typisches Beispiel für lange alphanumerische Variablen sind die für Datenbereitstellungen mit Hash-Algorithmen generierten Pseudonyme, die oftmals aus 64 Zeichen bestehen.

Darüber hinaus sollte zur Verringerung des Speicherbedarfs in großen Tabellen mit Millionen von Beobachtungen möglichst auch auf andere lange Textvariablen verzichtet werden. Beispielsweise spart es relevante Ressourcen, wenn statt einer genauen textlichen Bezeichnung einer ICD-10-Diagnose mit teils deutlich mehr als 100 Zeichen lediglich der eindeutig identifizierende 6-stellige ICD-10-Kode abgelegt ist. Um den Überblick zu behalten, lohnt es sich unter Umständen, mit Variablen- bzw. Wertelabel zu arbeiten. Mit Variablen- bzw. Wertelabels sind „Etiketten“ für Variablen oder für verschiedene Ausprägungen einer Variablen gemeint, die man häufig bei Statistikprogrammen sie SPSS oder SAS vergeben kann:

Variable	Geschlecht der Versicherten
Variablenname im Programm:	GESCH
Variablenlabel (z. B. in SAS, SPSS):	Geschlecht der Versicherten
Mögliche Ausprägungen der Variablen:	1 / 2 / 3
Wertelabel (z. B. in SAS, SPSS):	1 = „männlich“, 2 = „weiblich“, 3 = „divers“

Diese Etiketten ermöglichen es, längere Beschreibungen zu speichern, ohne direkt den Namen der Variablen oder deren Ausprägungen zu ändern. Allerdings können sie zu Schwierigkeiten bei der Übertragung von einem Analyseprogramm zu einem anderen führen.

4.3.2 Intervallaufbereitung

In den Routinedaten sind immer wieder auch Informationen über zeitliche Intervalle mit „Von“- und „Bis“-Datum gespeichert, z. B. zu Krankenhausaufenthalten, dem Versicherungsstatus, dem Pflegegrad

oder ähnlichem. Der Umgang mit Informationen aus diesen Zeitintervallen kann dabei aus unterschiedlichen Gründen erschwert sein.

Zum einen ist die Zuordnung der Zeitintervalle zu den Merkmalsausprägung nicht immer eindeutig, d. h. es kann aus unterschiedlichen Gründen überlappende Zeitintervalle mit unterschiedlichen Merkmalsausprägungen geben. Ein Beispiel sind mehrere überlappende Intervalle mit Informationen zum Versicherungsstatus, die z. B. bei gleichzeitiger Beschäftigung von Versicherten bei mehreren Arbeitgebern entstehen. Eine Summierung der ausgewiesenen Versicherungszeiten würde hier zu Fehlinformationen führen. Die Daten müssen daher zunächst so aufbereitet werden, dass die zeitliche Überlappung entfernt wird. Eine Möglichkeit, wie die Daten mithilfe des Programms SAS aufbereitet werden können, ist bei Grobe (2003) beschrieben.

Umgekehrt können auch Intervallfolgen ohne zeitliche Lücken und zugleich ohne Veränderung hinsichtlich der anderweitig erfassten Merkmalsausprägungen auftreten. Dies ist z. B. der Fall, wenn Versicherte ihren Wohnort innerhalb eines relativ kleinen Radius wechseln. In den Routinedaten bei der Krankenkasse werden dann zwei Zeilen für die alte und neue Adresse mit „von“- und „bis“-Datum angelegt. In den gelieferten Routinedaten für Forschungszwecke ist in der Regel dann nur noch die Postleitzahl auf drei Stellen gekürzt enthalten (nicht aber die genaue Adresse). Die (trunkierte) Wohnortangabe ist also in beiden Zeitintervallen identisch und die Intervalle können somit ohne Informationsverlust zusammengefasst werden. Eine Möglichkeit, wie die Daten mithilfe des Programms SAS aufbereitet werden können, ist bei Grobe (2018) beschrieben.

Darüber hinaus müssen mitunter Zeitintervalle aus unterschiedlichen Datentabellen (d. h. zu unterschiedlichen Inhalten) miteinander kombiniert werden. Beispielsweise für eine Auswertung der Krankschreibung bestimmter Berufsgruppen müssen personenbezogen dokumentierte Berufstätigkeitsintervalle mit individuellen Krankschreibungsintervallen kombiniert werden. Diese Zuordnung kann insofern komplex sein, als dass die Tabellen in der Regel eine m:n-Beziehung aufweisen. Dabei können z. B. mehrere Krankschreibungsintervalle in ein Berufstätigkeitsintervall fallen oder eine Krankschreibung erstreckt sich über mehr als ein Tätigkeitsintervall. Eine Möglichkeit, wie die Daten mithilfe des Programms SAS aufbereitet werden können, ist in Grobe (2005) beschrieben.

Zusammenfassung von Krankenhausfällen

Für die Vorhersage von Nachsorgebedarfen nach Krankenhausentlassung musste zunächst eine Zusammenlegung von Krankenhausfällen durchgeführt werden. Grund dafür ist, dass bei Verlegung von Patienten in ein anderes Krankenhaus u.U. neue Krankenhausfälle in der Abrechnungstabelle entstehen (s. Kapitel 1). Der Nachsorgebedarf von Patienten soll aber nach Entlassung aus dem stationären Kontext erfolgen, also z. B. bei einer Entlassung nach Hause oder in ein Pflegeheim. Krankenhausaufenthalte sollen also im Idealfall mit der initialen Aufnahme in ein Krankenhaus beginnen und mit der Entlassung aus dem (letzten) Krankenhaus nach Hause (oder in eine andere Einrichtung) enden.

Von den insgesamt 4.090.658 Krankenhausfällen, die in den gelieferten Datentabellen enthalten waren, konnten 3.655.748 abgeschlossene Krankenhausepisoden (i.S. der oben genannten Definition) identifiziert werden. Dabei wurden zusätzlich Fälle mit der Hauptdiagnose „Geburt“ ausgeschlossen, d. h. Krankenhausaufenthalte von Neugeborenen.

4.3.3 Umgang mit zensierten Daten

In den Versichertenstammdaten der Krankenkassen liegen tagesgenaue Informationen zu den Versicherungszeiten der Versicherten vor (s. Kapitel 1). Mithilfe dieser Information kann überprüft werden, ob die Versicherten im gewählten Beobachtungszeitraum überhaupt durchgängig bei der Krankenkasse, deren Daten vorliegen, versichert waren und ob somit durchgängig Daten erfasst werden konnten. Sind Personen nicht durchgängig im Beobachtungszeitraum versichert gewesen, muss dieser Umstand berücksichtigt werden. Dies kann dazu führen, dass diese Versicherten gänzlich von Analysen

ausgeschlossen werden oder als zensierte Fälle analysiert werden müssen. In der Regel ist es möglich und notwendig, die Versicherten mit unzureichenden Versicherungszeiten komplett von der Analyse auszuschließen. Problematisch kann ein solches Vorgehen vor allem dann sein, wenn sich Personen mit unvollständigen Daten systematisch vom restlichen Kollektiv unterscheiden. Tendenziell scheinen vor allem junge und gebildete Versicherte häufiger die Krankenkasse zu wechseln, was zu unvollständigen Versicherungszeiten führen kann (Hoffmann & Icks, 2012). Insbesondere bei langen Beobachtungszeiten kann es also dazu kommen, dass diese Versichertengruppe überproportional häufig ausgeschlossen wird und somit unterrepräsentiert ist. Umgekehrt kann aber der Einschluss von Personen mit unvollständigen Versicherungszeiten zur Unterschätzung von Ereignishäufigkeiten führen, da während fehlender Zeiten kein Ereignis beobachtet werden kann (sofern Ereignisse nicht aus anderen Daten und per Annahme imputiert werden). Der Umgang mit zensierten Daten muss also je nach Situation überlegt werden.

Als Beobachtungszeitraum wurde hier der Krankenhausaufenthalt gewählt (mit variabler Länge je Versicherten). Die Prädiktoren und Outcomes im Beispiel setzen eine gewisse Mindestversicherungszeit vor und nach diesem Krankenhausaufenthalt voraus. Der Prädiktor „Arzneimittelverordnungen“ gibt zum Beispiel an, ob der oder die Versicherte in den drei Monaten vor der Aufnahme ins Krankenhaus insgesamt sechs oder mehr unterschiedliche Arzneimittel verordnet bekommen hat. Der oder die Versicherte muss also in diesen drei Monaten bei der datenliefernden Krankenkasse versichert gewesen sein, um mögliche Verordnungen zu identifizieren. Versicherte, die in dieser Zeit nicht durchgängig versichert waren, wurden für die Analyse vollständig ausgeschlossen.

4.4 Spezifische Aufbereitung des Analysedatensatzes für konventionelle und KI-basierte Vorhersagen

Für die Entwicklung von konventionellen und KI-basierten Vorhersagemodellen müssen die Routinedaten so aufbereitet werden, dass ein Analysedatensatz entsteht, der für die Entwicklung der Modelle genutzt werden kann. Das heißt, es müssen alle Variablen/Features in einer Datentabelle zusammengetragen werden, die zur Vorhersage eines interessierenden Merkmals relevant sein könnten. Die Variablen/Features können direkt in den Routinedaten enthalten sein (z. B. das Geburtsjahr oder Geschlecht der Versicherten). Zumeist müssen sie aber durch Kombination verschiedener Informationen aus den Routinedaten erstellt werden (z. B. Anzahl Arztkontakte in einem bestimmten Zeitraum). Dies wird im Bereich des Maschinellen Lernens auch als Feature Engineering bezeichnet. Hierbei müssen ggf. mehrere Datenverarbeitungsschritte durchgeführt werden, bevor die gewünschte Information verfügbar ist.

4.4.1 Aufbereitungsschritte und Erstellung des Analysedatensatzes

Für einen ersten Analysedatensatz wurden Variablen ausgewählt, die bereits im vorangegangen Projekt USER (gefördert durch den Innovationsausschuss beim Gemeinsamen Bundesausschuss unter dem Kennzeichen 01NVF18010) im Rahmen einer hypothesengeleiteten Literaturrecherche und empirischen Analysen als vielversprechend für die Vorhersage des Nachsorgebedarfs identifiziert worden waren. Tabelle 4-3 enthält eine Liste der Variablen des Analysedatensatzes, sowie eine kurze Beschreibung der Verarbeitungsschritte zur Erstellung der Variablen. Zur Veranschaulichung ist in Tabelle 4-4 ein beispielhafter Auszug aus dem Analysedatensatz gezeigt.

Tabelle 4-3. Liste der Variablen/Features im Analysedatensatz

Variable	Beschreibung	Berechnung
Versid	Versicherten-ID, numerisch	Ersetzen des zufällig vergebenen, alphanumerischen Pseudonyms (das bei der Lieferung vergeben wurde) durch numerisches Pseudonym
Episode	ID der Krankenhausepisode pro Versicherten (da mehrere Krankenhausaufenthalte pro Person möglich sind), numerisch	Bei der Intervallaufbereitung von Krankenhausfällen vergeben
Aufnahmedatum	Initiales Aufnahmedatum ins Krankenhaus	Erstes Aufnahmedatum nach Intervallzusammenlegung
Entlassdatum	Endgültiges Entlassdatum aus dem Krankenhaus	Letztes Entlassdatum nach Intervallzusammenlegung
Geschlecht	Geschlecht (männlich/weiblich: 0/1-codiert)	Entsprechend dem Eintrag in den Stammdaten
Alter	Alter in Jahren zum Zeitpunkt der Aufnahme	Berechnet als Jahr der Aufnahme ins Krankenhaus minus Geburtsjahr
OUT_Mortalität	Outcome: Versterben innerhalb von 30 Tagen nach Entlassung (0/1-codiert)	Prüfen, ob Todesdatum (sofern vorhanden) zwischen Entlassdatum und Entlassdatum + 30 Tage liegt
OUT_Wiederaufnahme	Outcome: Ungeplante Wiederaufnahme ins Krankenhaus innerhalb von 30 Tagen nach Entlassung (0/1-codiert)	Aufnahmedatum ins Krankenhaus mit Aufnahmegrund = 7 („Wiederaufnahme wegen Komplikationen“) und innerhalb

Variable	Beschreibung	Berechnung
		von 30 Tagen nach Entlassung der vorherigen Krankenhausepisode
RISK_Multi-KH	Mehr als ein Krankenhausaufenthalt innerhalb von 6 Monaten vor der Aufnahme (0/1-codiert)	Krankenhausaufenthalte zwischen Aufnahmedatum und Aufnahmedatum – 6 Monate zählen
RISK_Long-KH	Mindestens ein Krankenhausaufenthalt mit Verweildauer >21 Tagen in den 365 Tagen vor Aufnahme (0/1-codiert)	Krankenhausaufenthalte zwischen Aufnahmedatum – 365 Tage und Aufnahmedatum identifizieren und auf Verweildauer (Entlassdatum-Aufnahmedatum) prüfen
RISK_Polymedikation	Mindestens 6 unterschiedliche Arzneimittelverordnungen innerhalb von 3 Monaten vor Aufnahme (0/1-codiert)	Identifizieren von Arzneimittelverordnungen zwischen Aufnahmedatum – 91 Tagen und Aufnahmedatum; unterschiedliche Wirkstoffe anhand von ATC-Codes zählen
RISK_Hilfsmittel	Mindestens eine Hilfsmittelverordnung in den 365 Tagen vor Aufnahme (0/1-codiert)	Prüfen der Hilfsmittelverordnungen zwischen Aufnahmedatum – 365 Tage und Aufnahmedatum
RISK_Pflegegrad	Pflegegrad zum Zeitpunkt der Aufnahme (von 0 bis 5)	Überprüfen, ob ein Pflegegradintervall (mit von- und bis-Datum) existiert, dass das Aufnahmedatum enthält und wenn ja, Pflegegrad extrahieren (sonst Pflegegrad = 0)
RISK_A00_A09 bis RISK_Z80_Z99	Vorhandensein min. einer ICD-Diagnose (ambulant oder stationär) innerhalb der jeweiligen dreistelligen ICD-Gruppe (insgesamt 241 verschiedene ICD-Gruppen: A00-A09, A15-A19, A20-A28 … usw. bis Z80-Z99, somit 241 unterschiedliche Variablen, jeweils 0/1-codiert)	Alle ambulanten und stationären Diagnosen im Zeitraum von Aufnahmedatum - 365 Tage bis Aufnahmedatum identifizieren und prüfen, ob die Anzahl Diagnosen in der jeweiligen Diagnosegruppe > 0 ist

Exemplarisch soll an dieser Stelle die Erstellung des letzten Merkmals, das Auftreten von ambulanten und stationären Diagnosen einer bestimmten ICD-Gruppe vor dem Krankenhausaufenthalt („Vorerkrankungen“), beschrieben werden. Ambulante und stationäre Diagnosen liegen in getrennten Tabellen vor. Für jeden Abrechnungsfall und Versicherten sind jeweils unbegrenzt viele Diagnose-Einträge (Zeilen) möglich. Um den Zeitpunkt des Auftretens einer Diagnose zu bestimmen, muss zunächst über die Fallnummer der zugehörige Abrechnungsfall zugeordnet werden. Da ambulante Fälle quartalsweise abgerechnet werden und deshalb mehrere Arztbesuche (bei dem gleichen Arzt/der gleichen Ärztin) im selben Quartal zu einem Fall zusammengefasst werden, kann das genaue Datum einer Diagnosestellung hier nicht eindeutig rekonstruiert werden. Wir nutzen das eingetragene Fall-End-Datum, um Diagnosen annähernd einem Zeitraum zuzuordnen. Ausgehend vom jeweiligen Aufnahmedatum jeder Krankenhausepisode selektieren wir alle ambulanten und stationären Diagnosen, die in den zurückliegenden 365 Tagen auftraten (bzw. bei denen das Fall-End-Datum der zugehörigen Abrechnungsfälle in diesem Zeitraum liegt) in Form einer Liste. An dieser Stelle sind auch alternative Kodierungen möglich. Wir haben für ein komplexeres Model zum Beispiel zusätzlich noch eine genauere Zuordnung der Diagnosen zu 3-Monats-Abschnitten durchgeführt. Generell bietet sich insbesondere bei den ambulanten Diagnosen auch eine Zuordnung zu Quartalen statt Tagen/Monaten an, um die Abrechnungsgegebenheiten und Unsicherheit bezüglich des genauen Zeitpunkts besser abzubilden. Zur Zuordnung der nun als Liste vorliegenden in der Regel 3- bis 5- stelligen ICD-Codes zu Diagnosegruppen oder -kapiteln wird eine entsprechende Übersetzungsdatei benötigt. Schließlich wird für jede ICD-Gruppe

bzw. jedes ICD-Kapitel eine Prädiktorvariable angelegt und die Häufigkeit des Auftretens des entsprechenden Codes in der Liste gezählt. Nachfolgend haben wir diese Zählvariable zu einer dichotomen Variable vereinfacht (0 oder >=1).

Tabelle 4-4. Beispielhafter Auszug aus der Datentabelle des KI-THRUST Analysedatensatzes

Versid	Episode	Aufnahmedatum	Entlassdatum	Geschlecht	Alter	OUT_Mortalität	OUT_Wieder-aufnahme	…	RISK_Z70 _Z76	RISK_Z80 _Z99
1	1	2018-06-12	2018-06-15	0	58.0	0	0		0	0
2	1	2018-04-21	2018-04-23	0	69.0	0	0		0	0
3	1	2018-04-11	2018-04-13	0	74.0	0	0		0	0
3	2	2018-06-26	2018-07-04	0	74.0	0	0		0	1
4	1	2018-04-25	2018-05-16	1	89.0	0	0		0	1
4	2	2018-10-22	2018-10-25	1	89.0	0	0		1	1
5	1	2018-09-28	2018-09-29	1	74.0	0	0		0	1
6	1	2018-02-18	2018-03-08	0	79.0	0	0		0	0
6	2	2018-10-06	2018-10-30	0	79.0	0	0		0	0
6	3	2018-11-15	2018-11-17	0	79.0	0	1		0	0
7	1	2018-01-12	2018-01-17	0	31.0	0	0		0	0
8	1	2018-01-20	2018-01-28	0	65.0	0	0		0	0
9	1	2018-03-01	2018-03-09	1	86.0	0	0		0	0
9	2	2018-05-21	2018-05-25	1	86.0	0	0		1	0
10	1	2018-08-20	2018-08-28	1	87.0	0	0		0	0
11	1	2018-09-11	2018-09-20	0	63.0	0	0		0	1

4.4.2 Dummy-Kodierung / One-Hot-Encoding

Viele Vorhersagemodelle, sowohl konventionelle als auch KI-basierte Verfahren, können nicht unmittelbar mit kategorialen Daten umgehen, d. h. mit Zuordnungen zu Kategorien oder Gruppen. Im einfachsten Fall handelt es sich um eine kategoriale Variable mit lediglich zwei Ausprägungen (z. B. Verstorben: ja / nein). Die Informationen können hier als 0/1-kodierte Indikatorvariable aufbereitet werden (z. B. 0 = „nicht verstorben“, 1 = „verstorben“), mit der die Vorhersagemodelle rechnen können. Hat eine kategoriale Variable aber mehr als zwei Ausprägungen, muss sie über eine *Dummy-Kodierung oder One-Hot-Encoding* in mehrere Indikatorvariablen umgewandelt werden. Beispiele für kategoriale Variablen mit mehr als zwei Ausprägungen in den Routinedaten sind z. B. die „Art der Inanspruchnahme“ für ambulante Leistungen (Originalschein, Vertreterschein, Notfallschein etc.) oder der „Entlassgrund“ bei Krankenhausbehandlungen (z. B. Behandlung regulär beendet, Behandlung gegen ärztlichen Rat beendet, Tod etc.) und üblicherweise auch das Geschlecht (männlich, weiblich, divers/sonstige). Für die Dummy-Kodierung dieser Variablen mit k Ausprägungen werden k-1 Dummy-Variablen erstellt. Bei einer als Referenzkategorie gewählten Ausprägung erhalten alle Dummy-Variablen den Wert „0“, womit auch diese Kategorie aus den k-1 Dummy-Variablen herleitbar ist. Eine solche Kodierung ermöglicht es, dass die Vorhersagemodelle Informationen darüber bekommen, welche Ausprägung bei einer Person bzw. einem Fall vorliegt. Tabelle 4-5. Dummy-Kodierung der kategorischen Variablen „Geschlecht“ zeigt ein einfaches Beispiel einer Dummy-Kodierung des Merkmals Geschlecht mit k = 3 Kategorien. Die Dummy-Kodierung wird i.d.R. bei Regressionsverfahren angewendet und wird inzwischen von den Regressionsprozeduren der gängigen Statistikprogramme automatisch umgesetzt, so dass die Dummy-Variablen nicht vorab manuell erstellt werden müssen.

Tabelle 4-5. Dummy-Kodierung der kategorischen Variablen „Geschlecht“

Ausprägung der Variablen	Beschreibung	D1	D2
M	Männlich (Referenz)	0	0
W	Weiblich	1	0
D	Divers	0	1

Eine vergleichbare Vorgehensweise wird im Bereich des maschinellen Lernens als One-Hot-Encoding bezeichnet. Hier werden bei k Merkmalsausprägungen auch k One-Hot-Variablen gebildet, wie in Tabelle 4-6 beispielhaft dargestellt.

Tabelle 4-6. One-Hot-Encoding der kategorischen Variablen „Geschlecht“

Ausprägung der Variablen	Beschreibung	V1	V2	V3
M	Männlich	1	0	0
W	Weiblich	0	1	0
D	Divers	0	0	1

Das One-Hot-Encoding gehört in der KI-Sprachwelt zum sogenannten Feature Engineering. Damit ist der Prozess gemeint, Daten so aufzubereiten, dass ein ML-Modell möglichst gut trainiert werden kann. Dazu zählen unter anderem das Bereinigen von Daten, die Auswahl relevanter Merkmale und der Umgang mit kategoriellen und fehlenden Werten. Darüber hinaus gibt es weitere Möglichkeiten, die Daten durch Normalisieren oder andere Transformationen so zu präparieren, dass die zugrunde liegen-

den Muster für das Modell besser zugänglich sind. Beispiele hierfür wären eine Dimensionalitätsreduktion durch Methoden wie Principal Component Analysis (PCA), Feature Scaling durch Standardisierung oder einfache Log-Transformationen.

Solche erweiterten Techniken wurden im Anwendungsbeispiel nicht durchgeführt, um eine Vergleichbarkeit der logistischen Regression mit den komplexeren ML-Methoden zu garantieren.

Quellen

Grobe, T. G. (2003). Aufarbeitung von überlappenden zu eindeutig abgegrenzten Zeitintervallen unter Beibehaltung der in Primärdaten enthaltenen Informationen. In C. Becker & H. Redlich (Eds.), *Data Mining und Statistik in Hochschule und Wirtschaft. Proceedings der 7. KSFE* (pp. 45-48). Aachen: Shaker Verlag.

Grobe, T. G. (2005). Zuordnung von Ereigsnisintervallen zu Bezugszeiträumen in größeren Datenbeständen. In E. Rödel & R.-H. Bödeker (Eds.), *SAS: Verbindung von Theorie und Praxis. Proceedings der 9. KSFE*. Aachen: Shaker Verlag.

Grobe, T. G. (2018). Universelles Makro zur Zusammenfassung von (lückenlos) dokumentierten Zeitintervallen. In C. Weiß, R. Minkenberg, & R. Muche (Eds.), *KSFE 2018. Proceedings der 22. Konferenz der SAS-Anwender in Forschung und Entwicklung (KSFE)* (pp. 45-56). Aachen: Shaker Verlag.

Hoffmann, F., & Icks, A. (2012). Structural differences between health insurance funds and their impact on health services research: Results from the Bertelsmann Health-Care Monitor. *Gesundheitswesen 74*(5), 291–297.

5 Umsetzung konventioneller und ML-basierter Modellberechnungen

Für das Projekt KI-THRUST wurden verschiedene Modelle für zwei verschiedene Outcomes (Mortalität und Ungeplante Wiederaufnahme) berechnet.
Als konventionelles Verfahren wurde ein logistisches Regressionsmodell auf der Basis von Trainingsdaten entwickelt. Im Bereich der ML-Methoden wurden Random Forest Modelle, Neuronale Netze und AdaBoost Modelle trainiert. Hierbei erfolgte die Hyperparameteroptimierung mit einer 5-fachen Kreuzvalidierung. Zur Auswahl der optimalen Hyperparameter wurde die Fläche unter der Precision-Recall-Kurve genutzt.
Anschließend wurde die Vorhersagegüte aller Modelle auf Basis der Testdaten bestimmt, um die Vorhersage der Modelle miteinander zu vergleichen.

In Abbildung 4 wird der Analyseplan schematisch dargestellt. So wird von den zwei Outcomes Mortalität und Wiederaufnahmen ausgegangen, welche prädiktiv durch die in Abschnitt 5.1.3 beschriebenen Modelle M1 bis M3 mit verschiedenen Komplexitätsgraden modelliert werden. Die grundsätzliche Aufbereitung der Routinedaten erfolgt anhand der in Kapitel 4 erläuterten Schritte, wobei das spezifische Vorgehen für das Projekt KI-THRUST in Abschnitt 5.1.1 beschrieben steht. Das folgende Datensplitting in Trainings- und Testdaten, beschrieben in Abschnitt 5.1.2, ermöglicht eine getrennte Entwicklung und Evaluation der Modellierungsverfahren. Im nächsten Schritt der Trainingsdaten-Pipeline befinden sich die verschiedenen ML-Verfahren und das klassische Vorgehen mithilfe von logistischer Regression, welche in Kapitel 2 beschrieben werden. Die Implementierung der Verfahren wird in den Abschnitten 5.2 und 5.3 erläutert. Für die ML-Verfahren Random Forest, AdaBoost und das Künstliche Neuronale Netz werden weitere Validierungs- und Optimierungsschritte vorgenommen, darunter eine Kreuzvalidierung (Abschnitt Kreuzvalidierung), Upsampling/Downsampling und das Implementieren einer gewichteten Fehlerfunktion aufgrund der Unbalanciertheit der Daten sowie eine Hyperparameteroptimierung per Gittersuche (Abschnitt 5.4). Die Evaluation der Ergebnisse der verschiedenen Verfahren mithilfe von Evaluations- und Erklärbarkeitsmetriken ist in Kapitel 7 beschrieben, während Fehlklassifikationen der Modelle in Kapitel 8 diskutiert werden.

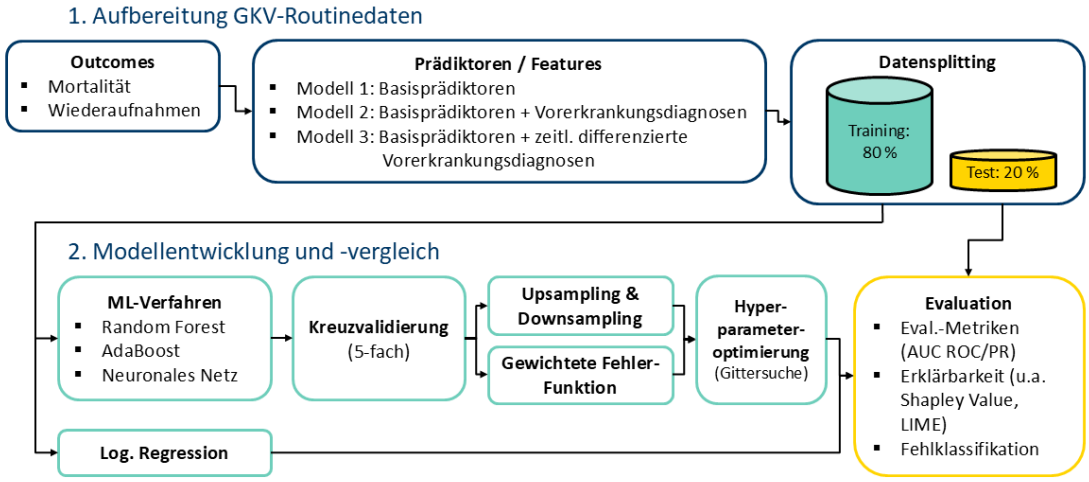

Abbildung 4 Studiendesign des methodischen Vorgehens bei KI-THRUST

Methoden der Modellentwicklung

Im folgenden Kapitel soll die Entwicklung von konventionellen und KI-basierten Vorhersagemodellen beispielhaft am Datensatz, der im Projekt KI-THRUST genutzt wurde, erläutert werden. Dabei werden zunächst Vorverarbeitungsschritte beschrieben, die sowohl für konventionelle als auch für ML-basierte Modelle durchgeführt wurden. Danach folgt die Beschreibung der logistischen Regression (als konventionelles statistisches Verfahren) und der ML-Vorhersagemodelle. Die Ergebnisse der beiden Verfahren werden dann im nachfolgenden Kapitel dargestellt und miteinander verglichen.

5.1 Vorverarbeitung

5.1.1 Datenaufbereitung

Um die Prognosemodelle zu berechnen, wurden zunächst die zeitlichen Intervalle der Krankenhausfälle aufbereitet. d. h. zwei oder mehr Krankenhausfälle wurden zusammengelegt, wenn sie durch eine Verlegung der Patienten entstanden waren. Im Anschluss kennzeichnete jeder Beginn eines stationären Falls die Aufnahme der Patienten ins Krankenhaus aus dem häuslichen Kontext und jedes Ende des Falls die Entlassung der Patienten nach Hause, in eine andere Einrichtung (z. B. Pflegeheim) oder möglicherweise das Versterben der Patienten. Im Anschluss wurden Krankenhausfälle mit der Hauptdiagnose „Geburt“ ausgeschlossen, sowie Fälle von Patienten, die zum Zeitpunkt der Aufnahme jünger als ein Jahr alt waren.

Als Beobachtungszeitraum für die Analysen wurde der Krankenhausaufenthalt gewählt (mit variabler Länge je Versicherten), sowie ein Vorbeobachtungszeitraum von 365 Tagen und ein Nachbeobachtungszeitraum von 30 Tagen. Versicherte, die in dieser Zeit nicht durchgängig versichert waren, wurden für die Analyse ausgeschlossen.

Zur Modellberechnung wurde der im vorangegangenen Kapitel vorgestellte Analysedatensatz genutzt. Basierend auf theoretisch fundierten Überlegungen und in den Vorprojekten[11] gewonnenen Erfahrungen sind in dem Datensatz sieben sogenannte Basisprädiktoren enthalten, die einen gesicherten Einfluss auf die gewählten Outcomes haben. Darüber hinaus sind Vorerkrankungsdiagnosen der Versicherten in den Daten enthalten.

5.1.2 Trennung von Trainings- und Testdaten

Vor der Berechnung der Modelle wurde der Datensatz in zwei Teile geteilt, einen Trainingsdatensatz und einen Testdatensatz. Für den Trainingsdatensatz wurden 80 % der Versicherten zufällig anhand ihrer entsprechenden ID-Variablen ausgewählt und alle Krankenhausepisoden dieser Versicherten wurden für den Trainingsdatensatz genutzt. Die Krankenhausaufenthalte der restlichen 20 % Versicherten wurden für den Testdatensatz genutzt. Da je Versicherten mehrere Krankenhausepisoden vorhanden sein können, führte diese Methode zu einer Teilung der Krankenhausepisoden im Verhältnis von 79,98 % Trainingsdaten und 20,02 % Testdaten. Der Trainingsdatensatz wurde genutzt, um die Modelle zu entwickeln und zu optimieren. Die Güte der Modelle wurde dann anhand der Testdaten gemessen und verglichen. Für das Modelltraining wurden Fälle mit Entlassungen im Jahr 2018 genutzt. Somit konnten die Prognosemodelle an Daten aus dem gleichen Jahr, sowie an Daten aus zukünftigen Jahren (2019 und 2020) getestet werden.

5.1.3 Berechnete Modelle

Im Rahmen der Analysen wurden jeweils Prognosemodelle für zwei Outcomes berechnet, **Mortalität** und **Ungeplante Wiederaufnahme** (jeweils dichotom kodiert als eingetreten/nicht eingetreten innerhalb von 30 Tagen nach Krankenhausentlassung, s. Abschnitt 4.1).

Für jedes Outcome wurden insgesamt drei Modelle berechnet, denen ein unterschiedlicher Satz an Variablen (Prädiktoren) zur Vorhersage angeboten wurden (s. Tabelle 5-1). Das **erste Modell (M1)** enthielt lediglich sieben Basisprädiktoren. Das **zweite Modell (M2)** enthielt neben den sieben Basisprä-

11 EMSE und USER, gefördert durch den Innovationsausschuss beim Gemeinsamen Bundesausschuss unter den Kennzeichen 01VSF16041 und 01NVF18010. Bericht USER: https://innovationsfonds.g-ba.de/beschluesse/user-umsetzung-eines-strukturierten-entlassmanagements-mit-routinedaten.203; Bericht EMSE: https://innovationsfonds.g-ba.de/beschluesse/emse-entwicklung-von-methoden-zur-nutzung-von-routinedaten-fuer-ein-sektorenuebergreifendes-entlassmanagement.3

diktoren zusätzlich die ambulanten und stationären Vorerkrankungsdiagnosen, die als 241 ICD-Gruppen aufgenommen wurden und somit jeweils kodierten, ob mindestens eine Diagnose aus der jeweiligen ICD-Gruppe in den 365 Tagen vor Aufnahme ins Krankenhaus vorlag (s. Abschnitt 4.4 im vorherigen Kapitel). Darüber hinaus wurde ein **drittes Modell (M3)** berechnet. Für dieses Modell wurden die Vorerkrankungsdiagnosen zeitlich stärker differenziert. Dazu wurden für jede ICD-Gruppe vier Variablen generiert, jeweils bezogen auf einen Drei-Monats-Zeitraum von Q1 (1.-3. Monat vor Aufnahme ins Krankenhaus) bis Q4 (9.-12. Monat vor Aufnahme ins Krankenhaus). So kodierten die Variablen RISK_A00_A09_Q1, RISK_A00_A09_Q2, RISK_A00_A09_Q3 und RISK_A00_A09_Q4 jeweils die Information, ob mindestens eine ICD-Diagnose (ambulant oder stationär) innerhalb der ICD-Gruppe A00 bis A09 in dem jeweiligen Zeitraum vor Aufnahme ins Krankenhaus vorhanden war. Im Gegensatz zu M2 bleibt damit die Information enthalten, ob eine Diagnose kurz vor KH-Aufnahme gestellt wurde oder bereits viele Monate vorher und ob die Diagnose (erst) einmalig gestellt wurde oder bereits über einen längeren Zeitraum besteht.

Tabelle 5-1. Auswahl der Prädiktorvariablen für die Prognosemodelle

Variable	Beschreibung	
Modell M1		
Alter	Alter in Jahren zum Zeitpunkt der Aufnahme	Basisprädiktoren
Geschlecht	Geschlecht (männlich/weiblich: 0/1-codiert)	
RISK_Multi-KH	Mehr als ein Krankenhausaufenthalt innerhalb von 6 Monaten vor der Aufnahme (0/1-codiert)	
RISK_Long-KH	Mindestens ein Krankenhausaufenthalt mit Verweildauer >21 Tagen in den 365 Tagen vor Aufnahme (0/1-codiert)	
RISK_Polymedikation	Mindestens 6 unterschiedliche Arzneimittelverordnungen innerhalb von 3 Monaten vor Aufnahme (0/1-codiert)	
RISK_Hilfsmittel	Mindestens eine Hilfsmittelverordnung in den 365 Tagen vor Aufnahme (0/1-codiert)	
RISK_Pflegegrad	Pflegegrad zum Zeitpunkt der Aufnahme (von 0 bis 5)	
Modell M2		
Basisprädiktoren	(s. Modell M1)	
RISK_A00_A09 bis RISK_Z80_Z99	Vorhandensein min. einer ICD-Diagnose (ambulant oder stationär) innerhalb der jeweiligen dreistelligen ICD-Gruppe (insgesamt 241 verschiedene ICD-Gruppen: A00-A09, A15-A19, A20-A28 … usw. bis Z80-Z99, somit 241 unterschiedliche Variablen, jeweils 0/1-codiert) in den 365 Tagen vor Aufnahme ins Krankenhaus	
Modell M3		
Basisprädiktoren	(s. Modell M1)	
RISK_A00_A09_Q1, RISK_A00_A09_Q2, RISK_A00_A09_Q3, … bis RISK_Z80_Z99_Q4	Vorhandensein min. einer ICD-Diagnose (ambulant oder stationär) der ICD-Gruppe im jeweiligen 3 Monatszeitraum von Q1 (1.-3. Monat vor KH-Aufnahme) bis Q4 (9.-12. Monat vor KH-Aufnahme); insgesamt 4 x 241 Variablen, jeweils 0/1-codiert)	

5.2 Regressionsanalysen

Als konventionelle statistische Verfahren wurden logistische Regressionsmodelle zur Vorhersage der Outcomes Mortalität und Ungeplante Wiederaufnahmen berechnet. Um die Modellgüte im Vergleich zu den ML-Verfahren zu bewerten, wurde ein Testdatensatz erstellt und vor den Modellberechnungen getrennt abgelegt (s. Abschnitt 5.1). Die Modellberechnung erfolgte auf Basis der Trainingsdaten.

Das **erste Modell (M1)** enthielt lediglich die sieben Basisprädiktoren, die im Einschlussverfahren in einem Block in das Modell aufgenommen wurden. Die Variable „Alter“ wurde dabei als kategoriale Variable in 5-Jahres-Altersgruppen aufgenommen, da anzunehmen ist, dass das Alter keinen linearen Zusammenhang mit den Outcomes hat. Der Zusammenhang mit Mortalität sollte beispielweise eher U-förmig sein, d. h. die Mortalität ist bei älteren Personen stark erhöht, sowie bei ganz jungen Personen, wohingegen sie im mittleren Alter am geringsten ist. Für kategoriale Variablen mit mehr als zwei Ausprägungen wurden für die Berechnung der logistischen Regression Dummy-Variablen erstellt (s. Kapitel 4.4.2). Als Referenzkategorie wurde für die Variable „Alter“ die Kategorie „60-64 Jahre“ und für die Variable Pflegegrad die Kategorie „0“, d. h. kein Pflegegrad, gewählt.

Das **zweite Modell (M2)** enthielt neben den sieben Basisprädiktoren zusätzlich die ambulanten und stationären Vorerkrankungsdiagnosen. Im Modell M2 wurden zunächst die Basisprädiktoren in einem ersten Block in das Modell aufgenommen und danach wurden im zweiten Block die Vorerkrankungsdiagnosen iterativ durch eine schrittweise Vorwärtsselektion (stepwise-forward-selection) eingeschlossen. Dabei wurden schrittweise nur die Diagnosen ins Modell eingeschlossen, die die größte, statistisch signifikante Verbesserung des Modells erbringen. Das Modell wurde iterativ erweitert um Variablen mit dem jeweils größten F-Wert, sofern der p-Wert unter 0,05 lag und bereits in der Gleichung enthaltene Variablen wurden ausgeschlossen, sobald der p-Wert über 0,1 stieg.

Im **dritten Modell (M3)** waren neben den Basisprädiktoren die zeitlich differenzierten Vorerkrankungsdiagnosen enthalten. Wie bei Modell M2 wurden im Modell M3 zunächst die Basisprädiktoren in einem ersten Block in das Modell aufgenommen und danach im zweiten Block die zeitlich differenzierten Vorerkrankungsdiagnosen (4 x 241 Variablen) iterativ durch eine stepwise-forward-selection eingeschlossen mit den gleichen Selektionskriterien wie bei Modell M2.

Die Berechnung der individuellen vorhergesagten Wahrscheinlichkeiten (P) für das Eintreten eines Outcomes P(Y = 1) erfolgte unter Berücksichtigung der modellierten Effektkoeffizienten (β) über folgende transformierte Regressionsgleichung:

$$P(Y = 1|X_i = x_i) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}$$

Zur Berechnung der logistischen Regressionsmodelle wurde die Software SAS 9.4 (SAS Institute Inc., Cary, USA) verwendet. Weitere Informationen und alternative Software zur Berechnung der logistischen Regressionen finden sich in Tabelle 5-2.

In SAS wurde die Prozedur „proc logistic“ benutzt, um die Modelle zu berechnen:

```
proc logistic data=TRAIN_DATA;
class RISK_GESCHL (Ref="0") RISK_ALTER (Ref="60") RISK_PFL_GR
(Ref="0") RISK_HILFSM (Ref="0") risk_polymed (Ref="0")
RISK_MULTI_KH (Ref="0") RISK_LONG_KH (Ref="0")/Param=REF;
Model_M1: model OUT_MORT (event='1')=RISK_GESCHL RISK_ALTER
RISK_PFL_GR RISK_HILFSM RISK_POLYMED_MULTI_KH RISK_LONG_KH;
```

```
score data=TEST_DATE fitstat out=OUTPUTFILE;
ods output ResponseProfile=M1_RP ModelANOVA=M1_AN
OddsRatios=M1_OR
Association=M1_AS;
run;
```

Für die Modellgütebestimmung und die Auswahl des Modells mit der besten Vorhersage wurden die Receiver-Operating-Characteristic (ROC-Kurve) und die dazugehörige Fläche unter der ROC-Kurve (engl. „area under the curve“, kurz: AUC-ROC) berechnet. Als zusätzliche Evaluationsmetrik wurde zudem die Precision-Recall-Kurve (PR-Kurve) und die Fläche unter der PR-Kurve (AUC-PR) berücksichtigt.

Tabelle 5-2. Software zur Berechnung einer logistischen Regression

Programm	Umsetzung der logistischen Regression	Weiterführende Information
SAS	Innerhalb der Prozedur „proc logistic“	https://documentation.sas.com/doc/en/statug/15.2/statug_logistic_syntax01.htm
SPSS	Prozedur „LOGISTIC REGRESSION“ oder im Menü unter dem Reiter „Analysieren“ > „Regression“ > „Binär logistisch…“	https://www.ibm.com/docs/en/spss-statistics/saas?topic=regression-logistic
STATA	Prozeduren „logistic“ (berichtet Odds Ratios) oder „logit“ (berichtet Koeffizienten)	https://www.stata.com/features/overview/logistic-regression/
R	Zum Beispiel über die Funktion „glm()“ mit dem Zusatz „family=binomial“	https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm
Python	Zum Beispiel im Package „scikit-learn“ unter dem Namen „LogisticRegression“	https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

5.3 ML-Verfahren

5.3.1 AdaBoost

Für den AdaBoost (Adaptive Boosting) Klassifikator haben wir die Standardimplementierung des Scikit-Learn Python Pakets verwendet (https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html). Wichtige Hyperparameter für AdaBoost sind:

- *n_estimators*: Beschreibt die Anzahl der schwachen Klassifikatoren, in unserem Fall Entscheidungsbäume, die trainiert werden sollen. Eine Erhöhung dieses Wertes kann die Genauigkeit verbessern, die Trainingszeit jedoch verlängern. In der Standardeinstellung sind 50 Entscheidungsbäume. Dies ist eine gute Balance zwischen Leistung und Rechenaufwand.
- *learning_rate*: Beschreibt einen Gewichtungsfaktor, der die Beiträge der einzelnen Klassifikatoren anpasst. Ein niedrigerer Wert, wie beispielsweise 0.1 kann zu einer besseren Generalisierung führen, während ein höherer Wert wie 1.0 ein schnelleres Lernen ermöglicht, jedoch auch das Risiko einer Überanpassung erhöhen kann.

Code Beispiel

```
# Importieren des AdaBoostclassifiers
from sklearn.ensemble import AdaBoostClassifier
# Importieren des DecisionTreeClassifiers
from sklearn.tree import DecisionTreeClassifier

# Laden des AdaBoost Modells mit gewählten Parametern
clf = AdaBoostClassifier(base_estimator=DecisionTreeClassi-
fier(), n_estimators=100, learning_rate=0.1)
# Trainieren des Modells auf den Daten X mit den Label y
clf.fit(X, y)
```

5.3.2 Random Forest

Für den Random Forest Klassifikator haben wir die Standardimplementierung des Scikit-Learn Python Pakets verwendet (https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html). Wichtige Hyperparameter sind:

- *n_estimators*: Beschreibt die Anzahl der Entscheidungsbäume im Random Forest. Eine höhere Anzahl kann die Genauigkeit erhöhen, erfordert jedoch mehr Rechenzeit. Standardmäßig werden 100 Entscheidungsbäume verwendet.
- *criterion*: Beschreibt die Funktion zur Messung der Qualität eines Splits. ’gini’ ist die Standardeinstellung.
- *max_depth*: Beschreibt die maximale Tiefe der einzelnen Entscheidungsbäume, was eine Überanpassung verhindern kann. Standardmäßig ist dieser Parameter auf ’None’ gesetzt, das heißt, dass die Bäume so tief wachsen, bis alle Blätter rein sind, oder weniger als ’min_samples_split’ Proben enthalten.

Code Beispiel

```
# Importieren des RandomForestClassifiers
from sklearn.ensemble import RandomForestClassifier

# Laden des Random Forest Modells
clf = RandomForestClassifier(max_depth=2, random_state=0)

# Trainieren des Random Forest Modells auf den Daten X mit dem
# Label y
clf.fit(X, y)
```

5.3.3 Künstliches Neuronales Netz

Das hier genutzte KNN ist ein Multilayer Perceptron (MLP), das in PyTorch implementiert wurde. MLPs sind leistungsstark in der Modellierung komplexer, nicht-linearer Beziehungen und können durch Hyperparameter-Optimierung stark verbessert werden. MLPs sind flexibel und anpassungsfähig, erfordern jedoch oft mehr Rechenressourcen und eine sorgfältige Hyperparameter-Optimierung. Das hier implementierte MLP besteht aus sechs voll verbundenen Schichten (Fully Connected Layers), die durch Dropout-Schichten getrennt sind. Diese Schichten helfen, eine Überanpassung (Overfitting) zu reduzieren, indem sie zufällig Neuronen während des Trainings deaktivieren. Jede Schicht verwendet die ReLU-Aktivierungsfunktion. Am Ende wurde eine Softmax-Aktivierung zur Klassifikation verwendet. Die Initialisierung der Gewichte erfolgt mittels Xavier-Uniform-Init und die Verzerrungen (Biases) werden auf null gesetzt, um das Modell effizient zu trainieren. Die wichtigsten Hyperparameter sind:

- *batch_size*: Beschreibt die Anzahl der Samples pro Trainingsbatch. Kleinere Batches führen zu schnellerer Aktualisierung der Gewichte und schnellerem Training, während größere Batches stabilere Schätzungen der Gradienten liefern. Typische Werte liegen zwischen 16 und 128, diese können jedoch je nach Datensatz stark variieren.
- *learning_rate*: Beschreibe die Lernrate des Optimierers. Ein niedrigerer Wert kann zu langsamem Training führen, während ein höherer Wert zu instabilem Training und Überanpassung führen kann. Typische Werte liegen im Bereich von 0.001 bis 0.1.

Code Beispiel

```
# Importieren des Pytorch Pakets
import torch
# Importieren des Pytorch.nn Sub-Paket
import torch.nn as nn s
# Importieren einiger Initialisierungsmethoden
from torch.nn.init import xavier_uniform_, zeros_

#Initialisieren des Neuronalen Netzes
class KIThrustMLP(nn.Module):
      def __init__(self, n_features):
            #Konstruktor des NN
            super(KIThrustMLP, self).__init__()
            #Dropout Layer
            self.dropout1 = nn.Dropout(0.2)
            #Linearer Layer
```

```
                self.fc1 = nn.Linear(n_features, 1000)
                self.dropout2 = nn.Dropout(0.2)
                self.fc2 = nn.Linear(1000, 1000)
                self.dropout3 = nn.Dropout(0.2)
                self.fc3 = nn.Linear(1000, 500)
                self.dropout4 = nn.Dropout(0.2)
                self.fc4 = nn.Linear(500, 100)
                self.dropout5 = nn.Dropout(0.2)
                self.fc5 = nn.Linear(100, 10)
                self.dropout6 = nn.Dropout(0.2)
                self.fc6 = nn.Linear(10, 2)
                #Softmax Outputlayer
                self.softmax = nn.Softmax(dim=1)

                # Gewichtung Layer fc1
                xavier_uniform_(self.fc1.weight)
                xavier_uniform_(self.fc2.weight)
                xavier_uniform_(self.fc3.weight)
                xavier_uniform_(self.fc4.weight)
                xavier_uniform_(self.fc5.weight)
                xavier_uniform_(self.fc6.weight)

                # Biases von Layer fc1 auf 0 setzen
                zeros_(self.fc1.bias)
                zeros_(self.fc2.bias)
                zeros_(self.fc3.bias)
                zeros_(self.fc4.bias)
                zeros_(self.fc5.bias)
                zeros_(self.fc6.bias)

    # Definieren des Forward passes des Neuronalen Netzes mit
    # Dropout, Aktivierungsfunktion ReLu und Softmax Funktion
    def forward(self, x):
                x = self.dropout1(x)
                x = torch.relu(self.fc1(x))
                x = self.dropout2(x)
                x = torch.relu(self.fc2(x))
                x = self.dropout3(x)
                x = torch.relu(self.fc3(x))
                x = self.dropout4(x)
                x = torch.relu(self.fc4(x))
                x = self.dropout5(x)
                x = torch.relu(self.fc5(x))
                x = self.dropout6(x)
                x = self.softmax(self.fc6(x))
    return x
```

5.4 Optimierung der ML-Modelle

5.4.1 Umgang mit unbalancierten Daten

Nach ersten Betrachtungen der Häufigkeitsverteilungen der Vorhersagevariablen liegt bei allen eine starke Unbalanciertheit vor, d. h. eine Klasse ist deutlich öfter vorhanden als die andere Klasse. Im Projekt KI-THRUST gilt das sowohl für das Outcome Mortalität (Nicht-Versterben kommt deutlich häufiger vor als Versterben), als auch für das Outcome Ungeplante Wiederaufnahmen (keine Wiederaufnahme kommt deutlich häufiger vor als eine Wiederaufnahme). Die Unbalanciertheit ist allerdings bei Mortalität deutlich höher als beim Outcome Ungeplante Wiederaufnahmen. Dies kann die Vorhersagekraft von überwachten Lernmethoden stark beeinträchtigen, da diese häufig nur lernen, die Klasse vorherzusagen, die häufiger vorkommt und somit beim Training häufiger vom Modell gesehen wird. Um diesen Effekt zu reduzieren, haben wir sowohl Upsampling und Downsampling (siehe Absatz 5.4.2 und 5.4.3) als auch die gewichtete Fehler-Funktion (siehe Abschnitt 2.4.2) implementiert. Vorläufige Testungen haben ergeben, dass ein Upsampling und Downsampling von 80 % den größten Effekt hat, sowie eine Gewichtung der Fehlerfunktion basierend auf der Größe der Klassen. Es gibt hier keine Standardwerte, auf die man sich verlassen kann, da diese Methoden abhängig von der Größe des Datensatzes, der Anzahl an Vorhersagevariablen und dem Ausmaß der Unbalanciertheit sind.

5.4.2 Umgang mit Unbalanciertheit: Upsampling

Für das Upsampling haben wir die Implementierung des SMOTEN Algorithmus aus dem imbalanced-learn Python package (https://imbalanced-learn.org/stable/index.html) genutzt, um die kleinere Klasse um 80 % zu vergrößern.

Code Beispiel

```
# Importieren des Numpy Pakets
import numpy as np
# Importieren des Counter Pakets
from collections import Counter
# Importieren des SMOTE Algorithmus
from imblearn.over_sampling import SMOTEN

# Erstellung eines zufälligen Datensatzen mit 3 Variablen
X = np.array(["Var1"] * 60 + ["Var2"] * 70 + ["Var3"] * 30,
# Erstellung eines binären Outputs für den Datensatz X
dtype=object).reshape(-1, 1)
# Größe der Klassen vorher: {1: 40, 0: 20}
y = np.array([0] * 20 + [1] * 40, dtype=np.int32)
# Initialisierung von SMOTE
sampler = SMOTEN(random_state=0)
# Anwendung von SMOTE zum Upsampling
X_res, y_res = sampler.fit_resample(X, y)
# Größe der Klassen nachher {1: 40, 0: 40}
```

5.4.3 Umgang mit Unbalanciertheit: Downsampling

Für das Downsampling haben wir zufällig 80 % der größeren Klasse entfernt. Umgesetzt haben wir dies, indem wir zufällig 20 % der größeren Klasse ausgewählt haben, welche wir für das Training behalten haben.

Code Beispiel

```
# Indizes der Klassen 0 (größere Klasse) und 1 (kleinere Klasse)
zero_idx = (y == 0).nonzero(as_tuple=True)[0].tolist()
one_idx = (y == 1).nonzero(as_tuple=True)[0].tolist()
# Berechne wie viele Datenpunkte von Klasse 0 20 % entsprechen
#(abgerundet)
k = np.floor(len(zero_idx)*(1-0.8/100))
# Wähle zufällig 20 % der Datenpunkte von Klasse 0 aus
keep = random.sample(zero_idx, k)
# Behalte alle Datenpunkte der Klasse 1
keep.extend(one_idx)
target_feature = target_feature[keep]
```

5.4.4 Umgang mit Unbalanciertheit: Gewichtete Fehler-Funktion

Die Gewichtung der Fehler-Funktion lässt sich bei allen Methoden des überwachten Lernens im scikit-learn Python package (https://scikit-learn.org/stable/index.html) durch den Zusatz class_weight = "balanced" im Modellaufruf umsetzen, außer für Modelle wie AdaBoost, welche eine Gewichtung einzelner Datenpunkte während des Trainings vornehmen, weshalb es nicht möglich ist ihre Fehler-Funktion zu gewichten. Für KNNs müssen die Gewichte für die Klassen per Hand berechnet werden. Das Gewicht für Klasse 0 ist z. B. eins minus die Anzahl der Datenpunkte von Klasse 1 geteilt durch die Anzahl aller Datenpunkte.

Code Beispiel

```
# Scikit-learn
model=RandomForestClassifier(class_weight="balanced")
# Pytorch
from torch import count_nonzero
weight_zero = 1-((len(y)-count_nonzero(y))/len(y))
weight_ones = 1-(count_nonzero(y)/len(y))
```

5.4.5 Gittersuche und Validierung für optimale Hyperparameter

Da die Vorhersagekraft der Modelle des überwachten Lernens stark von der Wahl der Hyperparameter abhängen kann, haben wir uns dafür entschieden eine Gittersuche sowie eine 5-fache Kreuzvalidierung zu implementieren (siehe Abschnitt 2.3.2). Für die Gittersuche haben wir die GridSearchCV-Funktion des scikit-learn Python Pakets (https://scikit-learn.org/stable/index.html) benutzt. Dadurch verhindern wir, dass wir durch die Gittersuche optimale Parameter auf dem gesamten Trainingsdatensatz finden, welche schlechter auf andere Datensätze generalisieren (Overfitting). Wegen der Unbalanciertheit unserer Daten haben wir uns dafür entschieden die Gittersuche nach optimalen Parametern suchen zu lassen, welche die Fläche unter der Precision-Recall-Kurve maximieren (siehe Abschnitt 3.5).

Tabelle 5-3 zeigt die getesteten Hyperparameter-Kombinationen für alle Modelle. Die optimale Hyperparameter-Kombination wird dann final ausgewählt, als die, die im Mittel über alle 5 Train-Test-Splits die höchste Fläche unter der Precision-Recall-Kurve erreicht hat.

Wir haben uns nach Betrachtung der Ergebnisse der Gittersuche für Model M1 (siehe Abschnitt 5.2) dafür entschieden, die Gittersuche bei Modell M2 und oder Modell M3 nur noch für die Modelle mit gewichteter Fehler-Funktion durchzuführen und nicht mehr für die Modelle mit Upsampling und Downsampling, da beide Methoden vergleichbare Verbesserungen hinsichtlich der Performance erbracht haben und es bei der gewichteten Fehler-Funktion keinen Informationsverlust gibt. Des Weiteren haben wir die Anzahl der Hyperparameter für Modelle M2 und M3 reduziert, da wir keine Verbesserungen feststellen konnten und die Laufzeit dadurch drastisch reduziert werden konnte (s. Tabelle 5-3).

Tabelle 5-3. Ausprägung der getesteten Hyperparameter in der Gittersuche

Verfahren	Hyperparameter*	Getesteter Umfang
Modell M1		
Ada Boost	Learning Rate	0.1, 0.5, 1, 1.5, 2, 2.5
	Number Estimators	5, 100, 150, 200, 250, 300, 400, 500
Random Forest	Criterion	gini, entropy, log_loss
	Maximum Depth	None, 100, 500, 1000
	Number Estimators	50, 100, 150, 200, 250, 300, 500, 1000, 5000, 10000
Neuronales Netz	Batch Size	5000, 10000, 15000
	Learning Rate	0.0001, 0.001, 0.01
Modell M2 und Modell M3		
Ada Boost	Learning Rate	0.1, 0.5, 1, 1.5
	Number Estimators	5, 100, 200, 300, 400, 500
Random Forest	Criterion	gini
	Maximum Depth	None
	Number Estimators	500, 1000
Neuronales Netz	Batch Size	5000, 10000, 15000
	Learning Rate	0.0001, 0.001, 0.01

*Für die Bezeichnung der Hyperparameter wurden die englischen Begriffe gewählt, da diese in der Regel von den entsprechenden Softwarepaketen (z. B. scikit learn in Python) genutzt werden.

5.5 Bewertung der Modellgüte mit AUC-ROC und AUC-PR

Für den abschließenden Vergleich der logistischen Regressionsmodelle und der besten ML-Modellen wurde die Receiver-Operating Characteristic (ROC-Kurve) und die Precision-Recall-Kurve (PR-Kurve), sowie die jeweiligen Flächen unter den Kurven (AUC-ROC und AUC-PR) betrachtet. Für die Bewertung der Modellgüte kann man bei der ROC-Kurve auf gängige Interpretationshilfen (z. B. nach Hosmer und Lemeshow, s. Kapitel 3.4) zurückgreifen, wobei ein Wert von 0,5 eine Vorhersage „wie zufällig“ und ein Wert von 1.0 als „perfekt“ kennzeichnet (alle real Betroffenen werden als solche vorhergesagt, ohne dass dabei eine Beobachtung mit der Vorhersage fälschlich als betroffen klassifiziert wird). Im Gegensatz dazu hängt die Bewertung der Fläche unter der PR-Kurve von der Wahrscheinlichkeit der positiven Klasse ab. Aufgrund der Unausgewogenheit unserer Outcomes im vorliegenden Datensatz hat ein Zufallsmodell für das Outcome Mortalität einen **AUC-PR = 0,01** und für das Outcome Ungeplante Wiederaufnahme einen **AUC-PR = 0,08**. Diese sind somit die unteren Schranken für die AUC-PR Werte. Auch AUC-PR-Werte liegen idealtypisch, ähnlich wie bei den AUC-ROC-Werten, nahe 1, wobei der Wert 1 bei einer perfekten Vorhersage resultieren würde (alle als betroffen klassifizierten Beobachtungen sind auch real betroffen, wobei zugleich auch alle real Betroffenen als solche erkannt werden).

6 Ergebnisse

Für die Darstellung der Ergebnisse wird zunächst die Datenstruktur und die Stichprobe beschrieben. Um die verschiedenen Modelle und ML-Verfahren miteinander vergleichen zu können, werden dann für die beiden Outcomes "Mortalität" und "Ungeplante Wiederaufnahme" und die Modelle 1-3 die Modellparameter (u.a. AUC-ROC/AUC-PR) anhand der Trainingsdaten berechnet. Für die ML-Verfahren werden folglich die Trainingsparameter dargestellt. Auf der Grundlage der ROC-/und PR-Kurve und den dazugehörigen, berechneten AUC-Werten werden die Modelle und Verfahren dann auf Basis der Testdaten miteinander verglichen. Schließlich erfolgt noch eine Gegenüberstellung der Rechenzeiten der verschiedenen Modelle.

6.1 Studienpopulation

Von den insgesamt 4.090.658 Krankenhausfällen, die in den gelieferten Datentabellen enthalten waren, konnten nach Datenaufbereitung 3.655.748 abgeschlossene Krankenhausepisoden (mit Entlassung nach Hause oder in eine Nicht-Krankenhaus-Einrichtung) identifiziert werden. Im Zuge dieses Aufbereitungsschrittes wurden zusätzlich 234.232 Fälle (5,7 %) mit der Hauptdiagnose „Geburt" ausgeschlossen, d. h. Krankenhausaufenthalte von Neugeborenen, sowie teilstationäre Fälle. Im Anschluss wurden 5.718 Fälle von Kindern im Alter von unter 1 Jahr ausgeschlossen (0,16 %) und 776.635 Fälle, weil die Versicherungszeiten für die Berechnung der Prädiktoren und Outcomes nicht ausreichend waren (21,2 %). Die Daten wurden dann im Verhältnis 80 % zu 20 % in Trainings- und Testdaten geteilt. Für das primäre Training der Modelle wurden Fälle mit Entlassung im Jahr 2018 genutzt. Hierfür stand ein Datensatz mit insgesamt 471.025 Fällen zur Verfügung (s. Abbildung 6-1). Die Testdaten umfassen 118.767 Fälle mit einer Krankenhausentlassung im Jahr 2018. Darüber hinaus wurden die trainierten Modelle auch mit den Daten zu Krankenhausentlassungen aus den Jahren 2019 und 2020 getestet, um die prognostische Güte an weiteren (künftigen) Jahresscheiben zu überprüfen.

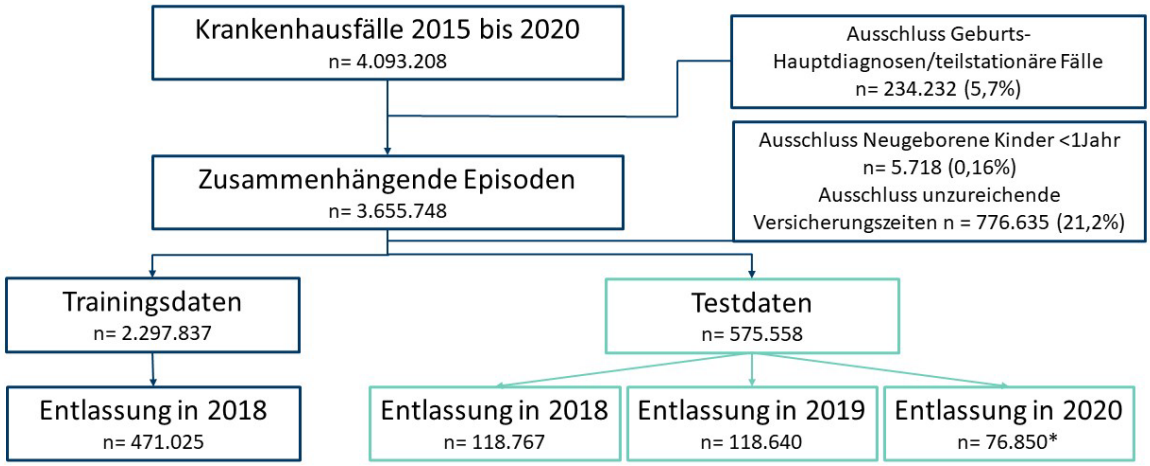

*Abbildung 6-1. Flowdiagramm der Studienpopulation (*für das Jahr 2020 wurden nur Daten bis zum 30.09. berücksichtigt)*

6.2 Beschreibung der Stichprobe

Die Versicherten im Trainingsdatensatz waren im Mittel 61,2 Jahre alt (Range von 1 bis 109 Jahre, SD = 22,85). Der Anteil an Männern im Datensatz lag bei 49,7 %, der Anteil Frauen bei 50,3 %. 82 Versicherte hatten den Geschlechtseintrag divers oder unbekannt. Diese Versicherten wurden aufgrund der geringen Anzahl vor der Erstellung des Trainingsdatensatzes ausgeschlossen. Eine detaillierte Beschreibung der Testdatensätze sowie Angaben zur Häufigkeit der untersuchten Outcomes und der Basisprädiktoren finden sich in Tabelle 6-1. Für die in der Tabelle beschriebenen Kennwerte zeigte sich kein signifikanter Unterschied zwischen dem Trainingsdatensatz (aus dem Jahr 2018) und dem Testdatensatz 2018 (alle $p > 0{,}17$), was der Intention bei einer zufälligen Aufteilung in Trainings- und Testdaten entspricht. Die verwendeten Testdatensätze aus den Jahren 2019 und 2020 enthielten demgegenüber signifikant mehr Versicherte mit Pflegegrad und Hilfsmittelverordnungen (alle $p < 0{,}01$). Im Testdatensatz 2020 kam außerdem das Outcome Mortalität signifikant häufiger vor ($p < 0{,}01$) und die Versicherten waren im Mittel 0,5 Jahre älter als im Trainingsdatensatz 2018 ($p < 0{,}01$), wobei sich die Unterschiede in einem Rahmen bewegen, der bei Daten zu Populationen aus unterschiedlichen Beobachtungsjahren zu erwarten ist.

Tabelle 6-1. Stichprobendeskription

Merkmal	Datensatz			
	Trainingsdaten 2018	Testdaten 2018	Testdaten 2019	Testdaten 2020
Stichprobengröße (n)	471.025	118.767	118.640	76.850*
Alter (M ± SD)	61,2 ± 22,9	61,2 ± 22,9	61,3 ± 22,9	61,7 ± 22,8
Altersgruppen in Jahren (%):				
Unter 20	6,7	6,8	6,7	6,7
20 bis 39	12,2	12,2	12,3	12,2
40 bis 59	20,3	20,3	20,0	19,5
60 bis 79	36,7	36,7	36,2	35,7
80 und älter	24,0	24,1	24,9	26,2
Geschlecht (%):				
Weiblich	50,3	50,5	50,2	50,2
Männlich	49,7	49,5	49,8	49,8
Pflegegrad (%):				
0	85,6	85,6	83,1	73,9
1	1,0	1,1	1,4	2,4
2	6,0	6,0	6,5	9,4
3	4,3	4,3	5,2	8,1
4	2,2	2,3	2,8	4,6
5	0,9	0,8	1,0	1,7
Outcome „Mortalität“ (%)	1,0	1,0	1,0	1,2
Outcome „Ungeplante Wiederaufnahmen“ (%)	8,2	8,2	8,3	7,9
Prädiktor „Polymedikation“ (%)	38,0	38,0	38,2	37,9
Prädiktor „Mehrfache KH-Aufenthalte“ (%)	16,0	16,2	16,4	15,6
Prädiktor „Langer KH-Aufenthalt“ (%)	8,8	8,9	9,0	8,7
Prädiktor „Hilfsmittelbedarf“ (%)	38,2	38,2	45,0	51,6

*für das Jahr 2020 wurden nur Daten bis zum 30.09. berücksichtigt

6.3 Berechnung der logistischen Regressionsmodelle

6.3.1 Outcome Mortalität

Das Modell M1 mit den Basisprädiktoren war signifikant besser als das Nullmodell, in das keine unabhängigen Variablen einfließt ($X^2(28) = 7.646{,}24$, $p < 0{,}001$). Mit dem zusätzlichen Einschluss der Vorerkrankungsdiagnosen in Modell M2 konnte das Regressionsmodell noch weiter verbessert werden. Die Erweiterung der Prädiktoren auf zeitlich differenziertere Diagnosegruppen im Modell M3 konnte das Regressionsmodell nicht weiter verbessern. Detaillierte Ergebnisse der Modelle, die auf Basis der Trainingsdaten aus dem Jahr 2018 berechnet wurden, sind in Tabelle 6-2 dargestellt.

Tabelle 6-2. Modellgüte der logistischen Regressionsmodelle für das Outcome Mortalität, basierend auf den Trainingsdaten aus dem Jahr 2018

Outcome: Mortalität	Modellparameter (auf Basis der Trainingsdaten 2018)			
Modell	**Cox-Snell R^2**	**AUC-ROC**	**AUC-PR**	**Anzahl eingeschlossener Variablen**
Modell M1	0,016	0,837**	0,048	7
Modell M2	0,023	0,892**	0,071	77
Modell M3	0,023	0,890**	0,074	86

*akzeptabel (0,7 ≤ AUC-ROC < 0,8) **ausgezeichnet (0,8 ≤ AUC-ROC < 0,9) ***hervorragend (0,9 ≤ AUC-ROC)

6.3.1 Outcome Ungeplante Wiederaufnahmen

Das Modell M1 mit den Basisprädiktoren war signifikant besser als das Nullmodell ($X^2(28) = 13.478{,}24$, $p < 0{,}001$). Mit dem Einschluss der Vorerkrankungsdiagnosen in Modell M2 konnte das Regressionsmodell noch weiter verbessert werden. Die Erweiterung der Prädiktoren auf zeitlich differenziertere Diagnosegruppen im Modell M3 konnte das Regressionsmodell nicht weiter verbessern. Detaillierte Ergebnisse der Modelle, die auf Basis der Trainingsdaten aus dem Jahr 2018 berechnet wurden, sind in Tabelle 6-3 dargestellt.

Tabelle 6-3. Modellgüte der logistischen Regressionsmodelle für das Outcome Ungeplante Wiederaufnahmen, basierend auf den Trainingsdaten aus dem Jahr 2018

Outcome: Ungeplante Wiederaufnahmen	Modellparameter (auf Basis der Trainingsdaten 2018)			
Modell	**Cox-Snell R^2**	**AUC-ROC**	**AUC-PR**	**Anzahl eingeschlossener Variablen**
Modell M1	0,028	0,680	0,149	7
Modell M2	0,035	0,698	0,164	108
Modell M3	0,036	0,698	0,166	153

*akzeptabel (0,7 ≤ AUC-ROC < 0,8) **ausgezeichnet (0,8 ≤ AUC-ROC < 0,9) ***hervorragend (0,9 ≤ AUC-ROC)

6.4 Training der ML-Verfahren

Um optimale Hyperparameter für die ML-Verfahren zu finden, haben wir uns dafür entschieden, eine Gittersuche sowie eine 5-fache Kreuzvalidierung zu implementieren (siehe Abschnitt 2.5.2). Als die optimale Hyperparameterkombination wird dann die Kombination final ausgewählt, die im Mittel über alle fünf Train-Test-Splits die höchste Fläche unter der Precision Recall Kurve erreicht hat.

Wir haben uns nach Betrachtung der Ergebnisse der Gittersuche für Model M1 (siehe Tabelle 6-4 und Tabelle 6-5) dafür entschieden, die Gittersuche bei Model M2 und oder Model M3 nur noch für die Modelle mit gewichteter Fehler-Funktion durchzuführen und nicht mehr für die Modelle mit Upsampling und Downsampling, da beide Methoden vergleichbare Performanceverbesserungen erbracht haben und es bei der gewichteten Fehler-Funktion keinen Informationsverlust gibt. Die Tabelle 6-4 und die Tabelle 6-5 zeigen für die unterschiedlichen Modelle und Outcomes die jeweils optimalen Hyperparametern. Für die Optimierung wurde die Fläche unter der PR-Kurve (AUC-PR) als Zielgröße festgelegt, da beide Outcomes vergleichsweise selten auftreten und folglich von tendenziell unbalancierten Daten auszugehen war.

Tabelle 6-4. Ausprägung der optimalen Hyperparameter für das Outcome Mortalität

		Ausprägung der optimalen Hyperparameter			
		Kein Up-/Downsampling		80 %-Up-/Downsampling	
Verfahren	**Hyperparameter**	**Gewichtet**	**Ungewichtet**	**Gewichtet**	**Ungewichtet**
Modell M1					
AdaBoost	Learning Rate	-	1,5	-	1,5
	Number Estimators	-	100	-	400
Random Forest	Criterion	gini	gini	gini	gini
	Maximum Depth	None	None	None	None
	Number Estimators	10.000	10.000	10.000	10.000
Neuronales Netz	Batch Size	15.000	5.000	5.000	15.000
	Learning Rate	0,0001	0,0001	0,001	0,01
Modell M2					
AdaBoost	Learning Rate	-	0,1	-	-
	Number Estimators	-	500	-	-
Random Forest	Criterion	gini	gini	-	-
	Maximum Depth	None	100	-	-
	Number Estimators	1.000	1.000	-	-
Neuronales Netz	Batch Size	5.000	15.000	-	-
	Learning Rate	0,0001	0,01	-	-
Modell M3					
AdaBoost	Learning Rate	-	0,1	-	-
	Number Estimators	-	500	-	-
Random Forest	Criterion	gini	gini	-	-
	Maximum Depth	None	100	-	-

		Ausprägung der optimalen Hyperparameter			
		Kein Up-/Downsampling		80 %-Up-/Downsampling	
Verfahren	Hyperparameter	Gewichtet	Ungewichtet	Gewichtet	Ungewichtet
Modell M1					
	Number Estimators	1.000	1.000	-	-
Neuronales Netz	Batch Size	5.000	10.000	-	-
	Learning Rate	0,0001	0,01	-	-

Tabelle 6-5. Ausprägung der optimalen Hyperparameter für das Outcome Ungeplante Wiederaufnahmen

		Ausprägung der optimalen Hyperparameter			
		Kein Up-/Downsampling		80 %-Up-/Downsampling	
Verfahren	Hyperparameter	Gewichtet	Ungewichtet	Gewichtet	Ungewichtet
Modell M1					
AdaBoost	Learning Rate	-	1	-	1,5
	Number Estimators	-	250	-	500
Random Forest	Criterion	entropy	gini	gini	gini
	Maximum Depth	None	None	None	None
	Number Estimators	10.000	10.000	10.000	10.000
Neuronales Netz	Batch Size	10.000	5.000	10.000	10.000
	Learning Rate	0,0001	0,01	0,0001	0,0001
Modell M2					
AdaBoost	Learning Rate	-	0,1	-	-
	Number Estimators	-	500	-	-
Random Forest	Criterion	gini	gini	-	-
	Maximum Depth	100	100	-	-
	Number Estimators	1.000	1.000	-	-
Neuronales Netz	Batch Size	5.000	15.000	-	-
	Learning Rate	0,0001	0,01	-	-
Modell M3					
AdaBoost	Learning Rate	-	0,1	-	-
	Number Estimators	-	500	-	-
Random Forest	Criterion	gini	gini	-	-
	Maximum Depth	100	100	-	-
	Number Estimators	1.000	1.000	-	-
Neuronales Netz	Batch Size	5.000	15.000	-	-
	Learning Rate	0,0001	0,01	-	-

6.4.1 Outcome Mortalität

Mit dem zusätzlichen Einschluss der Vorerkrankungsdiagnosen in Modell M2 konnten das Random Forest Modell, das AdaBoost-Modell und das KNN-Modell noch weiter verbessert werden. Die Erweiterung der Diagnosegruppen auf zeitlich differenziertere Diagnosegruppen im Modell M3 hat alle Modelle im Vergleich zum Modell M2 verschlechtert. Detaillierte Ergebnisse der Modelle, die auf Basis der Trainingsdaten aus dem Jahr 2018 berechnet wurden, sind in den folgenden Tabellen dargestellt.

Tabelle 6-6. Modellgüte der finalen ML-Modelle für das Outcome Mortalität, basierend auf den Trainingsdaten aus dem Jahr 2018

	Verfahren					
	AdaBoost		Neuronales Netz (KNN)		Random Forest	
Modell	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR
Modell M1	0,84**	0,05	0,81**	0,04	0,75*	0,04
Modell M2	0,89**	0,07	0,86**	0,05	0,88**	0,06
Modell M3	0,88**	0,07	0,84**	0,05	0,86**	0,05

*akzeptabel (0,7 ≤ AUC-ROC < 0,8) **ausgezeichnet (0,8 ≤ AUC-ROC < 0,9) ***hervorragend (0,9 ≤ AUC-ROC)

6.4.2 Outcome Ungeplante Wiederaufnahmen

Mit dem zusätzlichen Einschluss der Vorerkrankungsdiagnosen in Modell M2 konnten das Random Forest Modell und das AdaBoost-Modell noch weiter verbessert werden. Das KNN-Modell hingegen wurde verschlechtert. Die Erweiterung der Diagnosegruppen auf zeitlich differenziertere Diagnosegruppen im Modell M3 hatte keinen Einfluss auf das Random Forest Modell und das AdaBoost-Modell. Das KNN-Modell hingegen wurde verschlechtert. Detaillierte Ergebnisse der Modelle, die auf Basis der Trainingsdaten aus dem Jahr 2018 berechnet wurden, sind in den folgenden Tabellen dargestellt.

Tabelle 6-7. Modellgüte der finalen ML-Modelle für das Outcome Ungeplante Wiederaufnahmen, basierend auf den Trainingsdaten aus dem Jahr 2018

	Verfahren					
	AdaBoost		Neuronales Netz (KNN)		Random Forest	
Modell	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR
Modell M1	0,68	0,15	0,66	0,15	0,63	0,13
Modell M2	0,69	0,16	0,64	0,13	0,68	0,14
Modell M3	0,69	0,16	0,58	0,11	0,68	0,14

6.5 Vergleich logistische Regression und ML-Verfahren auf Basis der Testdaten

Nach der Entwicklung der logistischen Regressionsmodelle und der ML-Verfahren auf Basis der Trainingsdaten aus dem Jahr 2018 erfolgt der abschließende Modellvergleich mit Testdaten aus dem Jahr 2018. In den Vergleich gehen nur die Modelle ein, die sich im Zuge des Modelltrainings und der Hyperparameteroptimierung als die jeweils besten Varianten zur Vorhersage der Outcomes Mortalität und Ungeplante Wiederaufnahme erwiesen haben.

6.5.1 Outcome Mortalität

Die Bewertung der verschiedenen Verfahren und Datenmodelle erfolgt zunächst anhand der ROC-Kurven (siehe Abbildung 6-2) und den dazugehörigen AUC-ROC-Werten (siehe Tabelle 6-8). Vergleichen wir die drei Modellvarianten (M1 bis M3), so stellen wir für alle Vorhersageverfahren fest, dass die AUC-ROC-Werte von Modell 1 zu Modell 2 steigen, d. h. die Aufnahme der Vorerkrankungsdiagnose erhöhen die Vorhersagegüte gegenüber dem Modell 1, das lediglich auf den Basisprädiktoren (Alter, Geschlecht, etc.) basiert. Die zusätzliche, zeitliche Differenzierung der Vorerkrankungsdiagnosen im Modell 3 bringt hingegen keine weitere Verbesserung, sondern verringert sogar die Genauigkeit bei der Vorhersage der Mortalität. Somit lässt sich als erste Erkenntnis ableiten, dass das Modell 2 in Hinblick auf die ROC-Kurve bei allen Verfahren zur besten Vorhersage führt.

Beim Vergleich der vier Vorhersageverfahren fällt auf, dass alle AUC-ROC-Werte des Modells 2 in einem Bereich zwischen 0,8 und 0,9 liegen und somit gemäß etablierter Interpretationsempfehlungen (vgl. Tabelle 3-2) als „ausgezeichnet“ eingeordnet werden können. Das heißt, dass alle vier Verfahren in der Lage sind, die Mortalität auch mit den unbekannten Testdaten „ausgezeichnet“ vorherzusagen. Innerhalb dieses Wertebereichs lassen sich nur marginale Unterschiede zwischen den einzelnen Vorhersageverfahren erkennen, wobei AdaBoost (AUC-ROC = 0,889) und die logistische Regression (AUC-ROC = 0,888) etwas besser abschneiden, gefolgt von Random Forest (AUC-ROC = 0,878). Das neuronale Netz liegt mit einem AUC-ROC-Wert von 0,861 auf dem vierten Platz, was sich auch aus den ROC-Kurvenverläufen (siehe Abbildung 6-2) visuell ableiten lässt. Weiterhin fällt bei der Sichtung der Kurvenverläufe auf, dass das Random Forest-Verfahren beim Basismodell 1 eine relativ schlechte Performance liefert. Dieses Phänomen scheint aber im Modell 2 und Modell 3 durch die Aufnahme der Vorerkrankungsdiagnosen korrigiert zu werden.

Zieht man zusätzlich die PR-Kurven und die dazugehörigen Flächenwerte zur Modellbewertung heran, so ist auch hier festzustellen, dass bei allen Verfahren die AUC-PR-Werte von Modell 1 zu Modell 2 ansteigen. Ein Unterschied besteht allerdings beim Modell 3. Während bei Random Forest und dem neuronalen Netz die AUC-PR-Werte niedriger ausfallen (wie es auch bei AUC-ROC zu beobachten ist), bleibt der AUC-PR-Wert bei AdaBoost auf konstantem Niveau bzw. steigt im Fall der logistischen Regression sogar geringfügig von 0,068 auf 0,069 an (im Gegensatz zu AUC-ROC). Im Vergleich zu einem Nullmodell (also einer rein zufälligen Vorhersage) mit einem AUC-PR-Wert von 0,0098 ergeben sich bei allen Modellvorhersagen lediglich AUC-PR-Werte, die absolut nur um einen relativ kleinen Betrag höher als beim Nullmodell und damit insgesamt auf einem niedrigen Niveau liegen. Nach diesen Ergebnissen ist davon auszugehen, dass auch bei Personen, die gemäß der Modellvorhersagen sinngemäß “Höchstrisikogruppen” zugeordnet werden, nachfolgend nur ein kleiner Teil auch real von den vorhergesagten Ereignissen betroffen ist, was bei einer Nutzung der Vorhersagen bedacht werden muss.

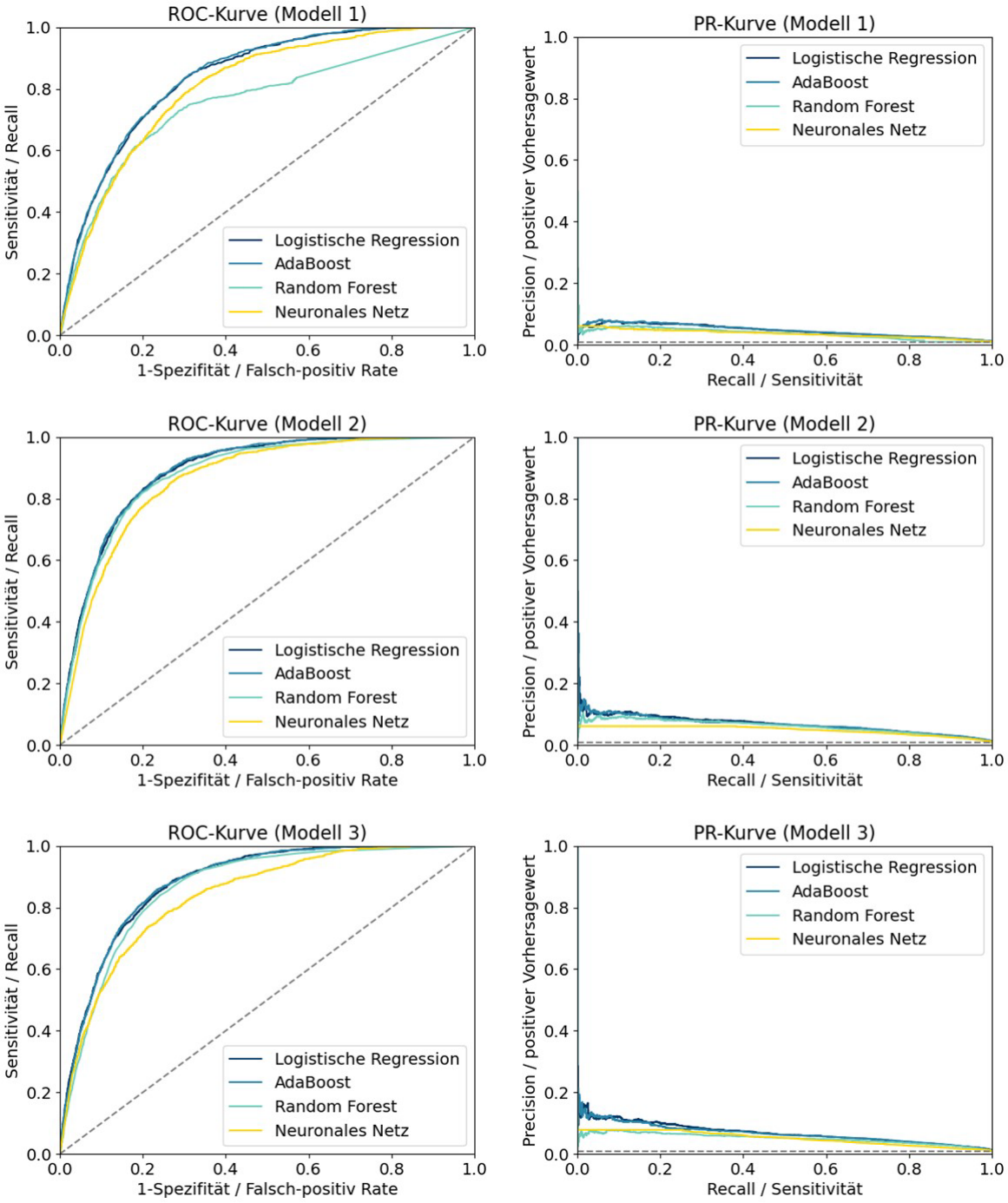

Abbildung 6-2. Receiver-Operating Characteristic (ROC-Kurve) und Precision-Recall-Kurve (PR-Kurve) für Verfahren auf Basis der Testdaten 2018 für das Outcome Mortalität

Tabelle 6-8. Modellgüte für das Outcome Mortalität, basierend auf den Testdaten aus dem Jahr 2018

Outcome: Mortalität	Ergebnisse					
	Modell 1		Modell 2		Modell 3	
Verfahren	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR
Logistische Regression	0,838**	0,046	0,888**	0,068	0,882**	0,069
AdaBoost	0,841**	0,048	0,889**	0,068	0,882**	0,068
Random Forest	0,755*	0,036	0,878**	0,061	0,864**	0,051
Neuronales Netz	0,807**	0,037	0,861**	0,048	0,838**	0,051

*akzeptabel (0,7 ≤ AUC-ROC < 0,8) **ausgezeichnet (0,8 ≤ AUC-ROC < 0,9) ***hervorragend (0,9 ≤ AUC-ROC)

6.5.2 Outcome Ungeplante Wiederaufnahmen

Wie bereits im vorangegangenen Abschnitt am Beispiel des Outcomes Mortalität demonstriert (siehe Kapitel 6.5.1), lassen sich auch die verschiedenen Verfahren, die zur Vorhersage des Outcomes Ungeplante Wiederaufnahme trainiert worden sind, anhand der ROC-Kurven (siehe Abbildung 6-3) und den dazugehörigen Flächenwerten (siehe Tabelle 6-9) bewerten und miteinander vergleichen. So fällt bei der Betrachtung der drei Modellvarianten (M1 bis M3) auf, dass nicht nur beim Outcome Mortalität, sondern auch beim Outcome Ungeplante Wiederaufnahme das Modell 2 bei allen Verfahren die höchsten AUC-ROC-Werte aufweist. Folglich ist es auch für die Vorhersage einer ungeplanten Wiederaufnahme ratsam, neben den Basisprädiktoren des Modells 1 auch die Vorerkrankungsdiagnosen für die Modellierung zu verwenden. Weiterhin können wir festhalten, dass alle Verfahren mit dem Modell 2 einen AUC-ROC-Wert erzielen, der unterhalb der Grenze von 0,7 liegt, die gemäß den Interpretationsempfehlungen (vgl. Tabelle 3-2) als „akzeptabel“ gilt. Somit lässt sich konstatieren, dass das poststationäre Ereignis einer ungeplanten Wiederaufnahme deutlich schwerer mit den hier berücksichtigten Krankenkassendaten zu prognostizieren ist als das Versterben.

Aus dem Vergleich der vier Vorhersageverfahren (jeweils M2) ergibt sich dasselbe Ranking wie bei der Mortalität. Auch bei der Wiederaufnahme liegt das ML-Verfahren AdaBoost (AUC-ROC = 0,694) knapp vor der logistischen Regression (AUC-ROC = 0,693). Mit etwas Abstand platziert sich das Random Forest-Verfahren (AUC-ROC = 0,681) auf dem dritten Rang. Das neuronale Netz (AUC-ROC = 0,638) liegt auf dem vierten Platz. Bei der Betrachtung des ROC-Kurvenverlaufs fällt zudem auf, dass das neuronale Netz noch am besten mit dem Basismodell 1 performt und mit dem Modell 2 und Modell 3 deutlich gegenüber den anderen drei Verfahren abfällt. Somit scheint das neuronale Netz deutlich weniger von der Aufnahme der Vorerkrankungsdiagnosen zu profitieren.

Vergleichen wir die ROC-Kurven mit den PR-Kurven (siehe Abbildung 6-3) und deren Flächenwerten (siehe Tabelle 6-9), so zeigen sich beim Outcome ungeplante Wiederaufnahme die AUC-Werte der ROC- und PR-Kurven weitgehend korreliert, d. h. der Vergleich der Vorhersageverfahren und Modellvarianten führt bei beiden Evaluationsmetriken zu denselben Einschätzungen hinsichtlich des Rankings. Dabei liegen die AUC-PR-Werte der getesteten Verfahren und Modelle in der Regel zwischen 0,13 und 0,16. Auch hier übersteigen die AUC-PR-Werte der Modellvorhersagen den entsprechenden Wert eines Nullmodells (AUC-PR = 0,08) nur um einen verhältnismäßig kleinen absoluten Betrag und bewegen sich insgesamt auf einem niedrigen Niveau. Entsprechend ist auch bei den Vorhersagen zur ungeplanten Wiederaufnahme davon auszugehen, dass, selbst in Gruppen, denen gemäß Modellvorhersagen relativ hohe Risiken zugewiesen werden, nachfolgend nur ein kleiner Teil der Personen auch real von ungeplanten Wiederaufnahmen betroffen ist.

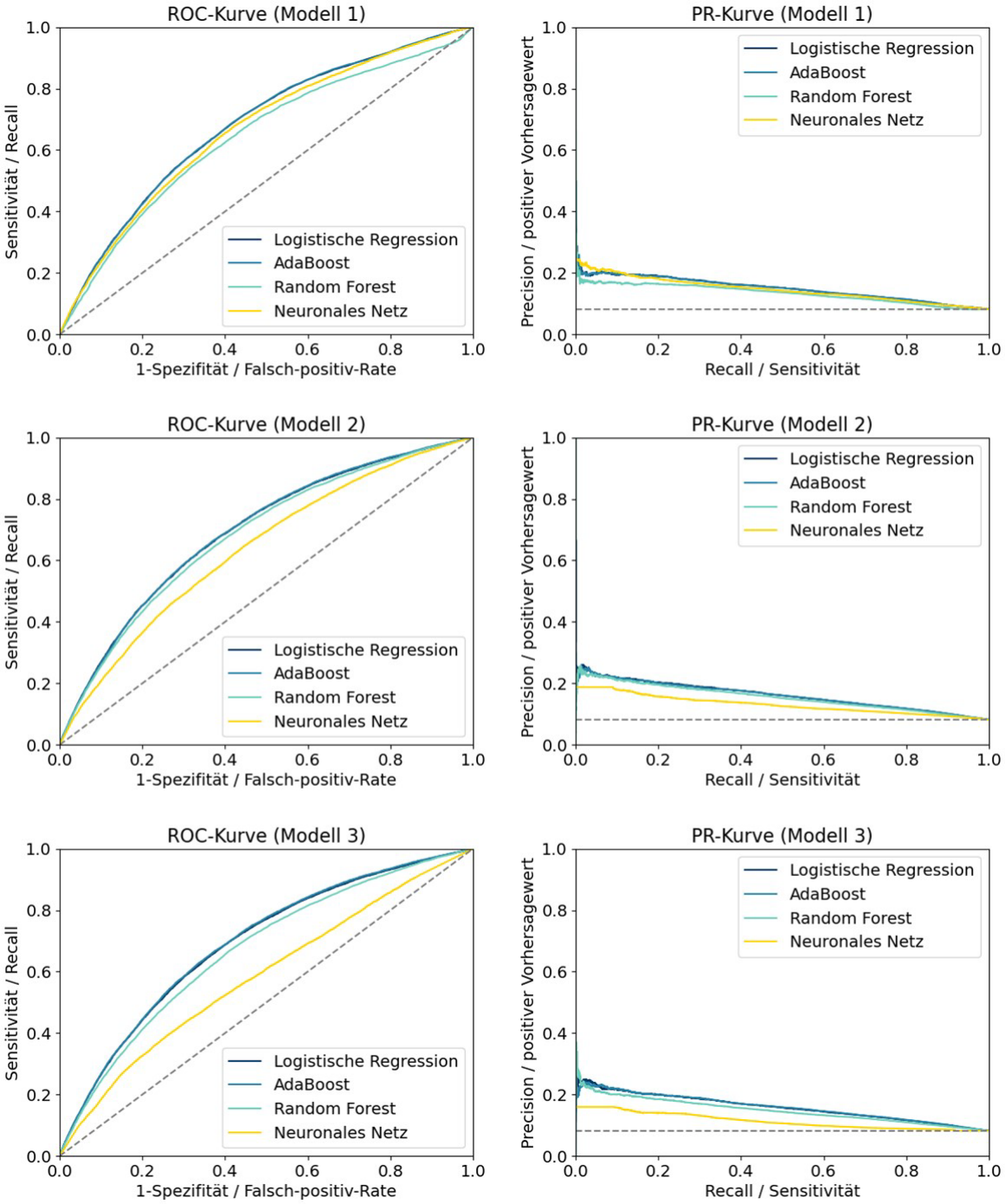

Abbildung 6-3. Receiver-Operating Characteristic (ROC-Kurve) und Precision-Recall-Kurve (PR-Kurve) für Verfahren auf Basis der Testdaten 2018 für das Outcome Ungeplante Wiederaufnahmen

Tabelle 6-9. Modellgüte für das Outcome Ungeplante Wiederaufnahmen, basierend auf den Testdaten aus dem Jahr 2018

Outcome: Ungeplante Wiederaufnahmen	Ergebnisse					
	Modell 1		Modell 2		Modell 3	
Verfahren	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR	AUC-ROC	AUC-PR
Logistische Regression	0,678	0,150	0,693	0,161	0,690	0,159
AdaBoost	0,678	0,150	0,694	0,160	0,693	0,159
Random Forest	0,641	0,134	0,681	0,154	0,669	0,148
Neuronales Netz	0,665	0,146	0,638	0,131	0,584	0,114

*akzeptabel (0,7 ≤ AUC-ROC < 0,8) **ausgezeichnet (0,8 ≤ AUC-ROC < 0,9) ***hervorragend (0,9 ≤ AUC-ROC)

6.6 Rechenzeiten der verschiedenen Verfahren

Die unterschiedlichen Verfahren haben teilweise stark unterschiedliche Berechnungszeiten. Um die Modelle vergleichen zu können, sind in Tabelle 6-10 die ungefähren Laufzeiten der finalen Modellberechnungen je Verfahren, Outcome und Modell (M1-M3) eingetragen. Dabei sind die Laufzeiten für die logistische Regression und die ML-Verfahren jedoch nur bedingt vergleichbar. Zum einen wurden die logistischen Regressionsmodelle auf einer anderen Hardware berechnet als die ML-Verfahren. Zum anderen fand bei der logistischen Regression keine Hyperparameter-Optimierung statt, so dass die angegebene Laufzeit die gesamte Rechenzeit für das Regressionsmodell widerspiegelt. Bei den ML-Verfahren hingegen waren die Laufzeiten zum Trainieren der Modelle teilweise deutlich länger als in der Tabelle angegeben (bis zu zwei Wochen pro Verfahren und Modell), da die Grid Search und Hyperparameteroptimierung ein mehrmaliges Durchlaufen der Modelle benötigt. Die in der Tabelle angegebenen Laufzeiten für das Modelltraining spiegeln aber nur ein einmaliges Training der ML-Modelle mit den jeweils optimalen Hyperparametern wider. Die genaue Spezifikation der Soft- und Hardware, die für die ML-Verfahren genutzt wurde, wird in Anhang 11.2 näher beschrieben.

Tabelle 6-10. Laufzeit der finalen Modellberechnungen (für ML-Verfahren mit den jeweils optimalen Hyperparametern)

	Laufzeitlänge					
	Modell 1		**Modell 2**		**Modell 3**	
Verfahren	**Training**	**Testung**	**Training**	**Testung**	**Training**	**Testung**
Outcome: Mortalität						
Logistische Regression*	Gesamt: 8 Min.		Gesamt: 3 Stunden		Gesamt: 19 Stunden	
AdaBoost	12 Sek.	2 Sek.	8 Min.	6 Sek.	18 Min.	13 Sek.
Random Forest	8 Min.	1 Min.	17 Min.	23 Sek.	40 Min.	32 Sek.
Neuronales Netz	3 Min.	1 Sek.	17 Min.	1 Sek.	29 Min.	1 Sek.
Outcome: Ungeplante Wiederaufnahmen						
Logistische Regression*	Gesamt: 6 Min.		Gesamt: 3,5 Stunden		Gesamt: 11 Stunden	
AdaBoost	1 Min.	2 Sek.	6 Min.	5 Sek.	18 Min.	13 Sek.
Random Forest	15 Min.	1 Min.	30 Min.	39 Sek.	76 Min.	1 Min.
Neuronales Netz	14 Min.	1 Sek.	16 Min.	1 Sek.	28 Min.	1 Sek.

*wichtiger Hinweis: Bei der log. Regression wurde eine andere Hardware verwendet als bei den ML-Verfahren

7 Erklärbarkeit und Nachvollziehbarkeit von Vorhersageergebnissen

Es gibt verschiedene Erklärbarkeitsmethoden im Bereich des Maschinellen Lernens, die dazu dienen, die wichtigsten Merkmale zu identifizieren, die ein Modell für den Vorhersageprozess nutzt. Grundsätzlich unterscheidet man zwischen modellintrinsischer Erklärbarkeit, auch Ante-hoc XAI genannt, und einer nachträglichen Erklärbarkeit, auch Post-hoc XAI genannt. Bei Ante-hoc Methoden, wie zum Beispiel die Wichtigkeiten der Merkmale (Feature Importance) von Random Forest und AdaBoost, können durch das ML-Modell selbst während des Modelltrainings berechnet werden. Post-hoc Methoden hingegen ermöglichen es, nach dem Modelltraining den Einfluss von Merkmalen auch bei ML-Verfahren zu berechnen, die selbst keine Erklärbarkeit bieten. Hierzu zählen zum einen modellunabhängige Verfahren wie LIME und Shapley Value Sampling sowie Verfahren wie Integrated Gradients, die spezifisch für Neuronale Netzwerke entwickelt wurden.
Im Rahmen von KI-THRUST wurden die genannten Erklärbarkeitsmethoden auf die unterschiedlichen ML-Methoden angewandt. Die Ergebnisse zeigen, dass die Methoden hilfreiche Hinweise auf relevante Merkmale geben können - beispielsweise bestimmte Vorerkrankungen - gleichzeitig aber auch methodische Grenzen und eine gewisse Instabilität aufweisen. Daher sind sie eher als unterstützendes Werkzeug zu verstehen und derzeit nur eingeschränkt für eine verlässliche inhaltliche Interpretation geeignet.

Wie in Kapitel 2 beschrieben, wurden im Rahmen von KI-THRUST sowohl interpretierbare, also intrinsisch erklärbare ML-Modelle (7.1), als auch Black-Box Methoden die eine zusätzliche Anwendung von Post-hoc Erklärbarkeitsmethoden bedürfen (7.2), angewendet.

7.1 Auswertung interpretierbarer Modelle

Zunächst wurden für die Analysen im KI-THRUST Projekt, mehrere intrinsisch interpretierbare ML-Methoden verwendet. Hierzu zählen insbesondere Random Forest und AdaBoost, für die die Wichtigkeiten der Merkmale (Feature Importance) durch das ML-Model selbst während des Modelltrainings berechnet werden können.

Wie in Kapitel 5.3.1 und Kapitel 5.3.2 beschrieben, werden die Merkmalswichtigkeiten auf Basis der Reduktion der Gini-Impurity (oder einer anderen Bewertungsmetrik wie Entropie) berechnet, die durch die Verzweigungen in den Entscheidungsbäumen erzielt werden. Der Algorithmus betrachtet, wie stark ein Split mit einem bestimmten Feature die Homogenität der Daten (z. B. in Bezug auf Klassen) verbessert. Diese Verbesserung wird summiert und über alle im Modell integrierten Entscheidungsbäume gemittelt. Die Werte geben an, wie viel jedes Feature relativ zu den anderen zur Vorhersage beiträgt. Ein Feature mit einer Importance von 0 hat keinen Beitrag geleistet. Ein Feature mit einer höheren Importance hat einen größeren Einfluss auf die Klassifikation.

Es gibt einen bekannten Bias zu numerischen und hochdimensionalen Features: Random Forests und AdaBoost können Features bevorzugen, die viele mögliche Verzweigungspunkte (z. B. kontinuierliche Werte) oder viele Kategorien (z. B. bei kategorialen Daten) haben. Die Feature Importances geben eine globale Übersicht über die Bedeutung eines Features, können aber wichtige lokale Unterschiede übersehen (Strobl et al., 2007; Neumann et al., 2017).

7.2 Anwendung der Erklärbarkeitsmethoden

Wie in Kapitel 2.7 beschrieben, ermöglichen es Post-hoc Erklärbarkeitsmethoden, nach dem Modelltraining den Einfluss von Merkmalen auch bei ML-Verfahren zu berechnen, die selbst keine Erklärbarkeit bieten.

Das Feld der Erklärbarkeitsmethoden entwickelt sich stets weiter, Methoden werden stets weiterentwickelt und neu entworfen, wodurch ein Konsens darüber, welche Methode besser auf spezifischen Daten funktioniert, bisher umstritten ist.

Um dennoch das Spektrum an unterschiedlichen Erklärbarkeitsmethoden abzudecken und vergleichen zu können, wurde jeweils eine Methode aus einer der Methodenkategorien gewählt und evaluiert, inwiefern sie sich unterscheiden oder übereinstimmen.

Zunächst wurde eine modellspezifische Methode namens Integrated Gradients verwendet, die speziell für Neuronale Netzwerke entwickelt wurde. Des Weiteren wurden zwei modellunabhängige Verfahren gewählt: Zum einen LIME, welche zu den Surrogate Methoden gehört, zum anderen das Shapley Value Sampling, welches eine Pertubationsmethode ist. Um eine Vergleichbarkeit der Methoden zu gewährleisten, wurden die Wichtigkeiten anschließend normalisiert.

7.2.1 Integrated Gradients

Integrated Gradients (IG) ist eine Methode der erklärbaren Künstlichen Intelligenz, die dabei hilft zu verstehen, wie ein ML-Modell eine Entscheidung trifft. Gerade bei komplexen Modellen, die oft als „Black Box“ angesehen werden, ist es schwierig nachzuvollziehen, warum bestimmte Vorhersagen gemacht werden. IG schafft hier Abhilfe, indem es den Einfluss einzelner Merkmale, also der Informationen, die dem Modell zur Verfügung stehen, quantifiziert.

IG funktioniert, indem es die Modellvorhersage entlang eines „Weges“ von einem neutralen Ausgangswert (Baseline) zur tatsächlichen Eingabe berechnet. Die Baseline kann als eine Art „Referenzwert“ angesehen werden, wie etwa ein „Standartpatient“ ohne Risikofaktoren. Durch das Vergleichen der Vorhersage für die Baseline und der Vorhersage für den tatsächlichen Patienten werden schrittweise die Unterschiede aufgezeigt, die jedes Merkmal auf die Entscheidung ausübt.

Der mathematische Prozess läuft folgendermaßen ab: Für jedes Merkmal, zum Beispiel das Alter, wird ein IG-Wert berechnet. Dieser gibt an, wie stark sich die Vorhersage ändert, wenn man vom Baseline-Wert für das Alter zum tatsächlichen Alter der Person geht. Die Berechnung dieses IG-Wertes erfolgt durch die Formel:

$$IG_i(x) = (x_i - x_i') \cdot \int_{a=0}^{1} \frac{\partial f(x' + a \cdot (x - x'))}{\partial x_i} da$$

Hierbei steht:

- f Für die Vorhersagefunktion des Modells,
- x Für die aktuelle Eingabe des Patienten, dessen Einfluss untersucht wird
- x' Für den Baseline Wert
- a Für einen Skalenfaktor, der von 0 bis 1 reicht.

Mit dieser Formel integriert IG den Einfluss des Merkmals „Alter“ auf die Modellvorhersage entlang des gesamten Wegs vom Baseline-Wert zum tatsächlichen Wert des Merkmals. Die Berechnung wird

für jedes Merkmal in den Krankenkassendaten durchgeführt. So erfahren wir beispielsweise, ob das Alter stark zur Entscheidung des Modells beigetragen hat.

Am Ende liefert IG für jedes Merkmal einen Wert, der angibt, wie relevant dieses Merkmal für die Modellentscheidung war. Solche Informationen können sehr nützlich sein, da sie Einblicke bieten, welche Faktoren die Risikoberechnung für bestimmte Outcomes besonders stark beeinflussen.

7.2.2 LIME (Local Interpretable Model-agnostic Explanations)

LIME ist eine Erklärungsmethodik für maschinelle Lernmodelle, die es ermöglicht, den Einfluss einzelner Merkmale auf eine bestimmte Modellentscheidung zu analysieren. Dabei ist LIME „modellagnostisch“, was bedeutet, dass es unabhängig vom zugrunde liegenden Modelltyp funktioniert und auf eine Vielzahl von Modellen anwendbar ist. Diese Flexibilität macht LIME besonders nützlich im Bereich der Krankenversicherungsanalyse, wo oft komplexe und intransparente Modelle zur Risikoschätzung oder Diagnosevorhersage verwendet werden (Ribeiro et al., 2016).

LIME bietet eine Lösung, indem es ein lokales, interpretierbares Modell erstellt, das erklärt, wie die verschiedenen Merkmale bei einer bestimmten Modellvorhersage zusammenwirken.

LIME geht davon aus, dass die globale Modellstruktur komplex sein kann, die Einflussfaktoren für eine einzelne Vorhersage aber durch ein vereinfachtes lineares Modell verständlich gemacht werden können. Die Methode funktioniert wie folgt:

- Generierung benachbarter Datenpunkte: LIME erzeugt synthetische, leicht veränderte Datenpunkte in der Umgebung des zu erklärenden Datenpunktes, indem die Werte der Merkmale minimal variiert werden. Für Krankenkassendaten könnte dies z. B. bedeuten, dass das Alter um einige Jahre erhöht oder verringert wird oder dass die Anzahl der diagnostizierten Vorerkrankungen verändert wird.
- Modellprognosen für die synthetischen Daten: Anschließend wird für jeden der synthetischen Datenpunkte eine Vorhersage aus dem Originalmodell generiert. Dies ermöglicht es, die Sensitivität des Modells auf Änderungen einzelner Merkmale zu messen. Eine starke Reaktion des Modells auf eine Änderung des Alters würde beispielsweise darauf hindeuten, dass das Alter eine zentrale Rolle für diese spezifische Vorhersage spielt.
- Anpassung eines lokalen Modells: Die synthetischen Datenpunkte und ihre Vorhersagen werden verwendet, um ein einfaches Modell (z. B. eine lineare Regression) zu erstellen, dass die Entscheidungslogik des ursprünglichen Modells für den spezifischen Datenpunkt lokal approximiert.
- Interpretation der Merkmalsgewichte: Dieses lokale Modell liefert Gewichte für jedes Merkmal, die dessen Einfluss auf die spezifische Vorhersage quantifizieren. Ein hohes Gewicht für „Anzahl der Vorerkrankungen“ bedeutet, dass dieser Faktor stark zur Risikoschätzung beiträgt. Dadurch wird sichtbar, welche Merkmale das Modell bei der Risikoschätzung am stärksten berücksichtigt hat.

LIME eignet sich unter anderem sehr gut für Krankenkassendaten, da es erlaubt, den Beitrag einzelner Gesundheitsmerkmale zur Entscheidung eines Modells nachvollziehbar darzustellen, ohne die gesamte Modellstruktur offen zu legen. Gerade in der Krankenversicherung, wo Entscheidungstransparenz von hoher Relevanz ist, bietet LIME damit Einblicke in die Funktionsweise komplexer Prognosemodelle. Es ermöglicht medizinischen Experten, die Grundlage von Modellentscheidungen für konkrete Fälle nachzuvollziehen und damit das Vertrauen in die Anwendung solcher Modelle zu stärken.

Zusammenfassend lässt sich sagen, dass LIME durch die Generierung eines lokal interpretierten Modells Einblicke in die Beitragsstärke einzelner Merkmale ermöglicht und damit komplexe, auf Krankenkassendaten basierende Modelle transparent und nachvollziehbar macht.

7.2.3 Shapley Values

Shapley-Werte sind eine spieltheoretische Methode, die verwendet wird, um den Beitrag einzelner Merkmale zur Vorhersage eines Modells zu erklären. Sie sind besonders wertvoll bei der Analyse komplexer Modelle, da sie eine faire und mathematisch fundierte Verteilung der Verantwortung für die Vorhersage auf die verschiedenen Eingabemerkmale ermöglichen. Bei Krankenkassendaten, die typischerweise Informationen wie Alter, Geschlecht und Diagnosehistorie enthalten, können Shapley-Werte helfen, den Einfluss jedes Merkmals auf eine spezifische Risikoabschätzung des Modells transparent darzustellen.

Die Berechnung basiert auf der Messung des „marginalen Beitrags" jedes Merkmals in allen möglichen Kombinationen der anderen Merkmale. Für ein Merkmal x_i wird geprüft, wie sich die Modellvorhersage ändert, wenn x_i zu verschiedenen Kombinationen anderer Merkmale hinzugefügt wird. Auf diese Weise wird der Einfluss des Merkmals x_i analysiert, wenn es zu verschiedenen Konstellationen beiträgt (Shapley, 2010).

7.2.4 Normalisierung

Aufgrund der unterschiedlichen Wertebereiche der Erklärbarkeitsmethoden war es notwendig diese zu standardisieren, daher wurde folgende Rangnormalisierung angewendet, um einen vergleichbaren Wertebereich zu erlangen:

Sei $1 \leq n, m \in \mathbb{N}, D \subseteq \mathbb{R}^{nxm}$ eine Teilmenge von Daten und $X \subseteq D$ eine Teilmenge von Testdaten. Für eine Testprobe $x \in X$ und eine feste XAI-Methode bezeichne $x_r \in \mathbb{R}^{nxm}$ die methodenspezifischen Relevanzzuweisungen für die Probe x. Die rangnormalisierte Relevanzzuweisung von x wird definiert als

$$x_r^{norm} = \frac{\text{rank}(\text{abs}(x_r))}{n \cdot m} \cdot \text{sign}(x_r)$$

Wobei die abs()-Funktion und die sign()-Funktion komponentenweise angewandt werden, und die rank-Funktion wie folgt definiert ist:

Sei $1 \leq l \in \mathbb{N}$. Gegeben eine Menge von Zahlen $x_1, \ldots, x_l$ mit den zugehörigen Ordnungsstatistiken $x_{(1)} \leq \cdots \leq x_{(l)}$ bezeichnet der Index (i) in Klammern die i-te Ordnungsstatistik der entsprechenden Zahl. Für alle nicht mehrfach auftretenden Zahlen gibt i ihren Rang an, während für alle mehrfach auftretenden Zahlen der Rang als arithmetisches Mittel ihrer entsprechenden i-ten Ordnungsstatistik berechnet wird.

7.3 Ergebnisse zur Erklärbarkeit

Im folgenden Kapitel werden die Ergebnisse für die verschiedenen Modelle des Maschinellen Lernens dargestellt. Diese Ergebnisse haben wir nur für das Modell 2 berechnet, da dieses die beste Performance hatte. Hier wollen wir kurz auf die Interpretation dieser Ergebnisse eingehen. Für das AdaBoost-Modell und Random Forest-Modell, welche interpretierbare Modelle sind, werden die globalen Feature Importance Werte dargestellt. Bei diesen handelt es sich um positive Zahlen, die keine obere Schranke haben. Je größer die Feature Importance für eine Variable ist, desto wichtiger war sie für die Vorhersage, wobei ein Wert von Null bedeutet, dass die Variable keinen Einfluss hatte. Des Weiteren handelt es sich hierbei um globale Werte, d. h. Sie berücksichtigen die gesamten Trainingsdaten.

Im Gegensatz dazu stehen die lokalen XAI-Relevanzzuweisungen der Methoden Integrated Gradients, Shapley Value Sampling und LIME. Diese benutzen zwar auch das Neuronale Netz, welches auf den gesamten Trainingsdaten trainiert wurde, jedoch werden die Relevanzzuweisungen für jeden Patienten in den Testdaten erzeugt (lokal). Nach der oben beschriebenen Normalisierung befinden sich diese in einem Wertebereich von [-1,1], wobei negative Relevanzzuweisungen bedeuten, dass diese Variable gegen die Vorhersage der selektierten Klasse spricht. Eine positive Relevanzzuweisung spricht hingegen für die Vorhersage der selektierten Klasse. Eine Relevanzzuweisung von Null bedeutet, dass diese Variable keinen Einfluss hatte. Um einen globalen Effekt zu simulieren, haben wir für jede Variable die Mediane über alle Relevanzzuweisungen der richtig klassifizierten Testdaten der positiven Klasse dargestellt.

7.3.1 Outcome Mortalität (Model 2)

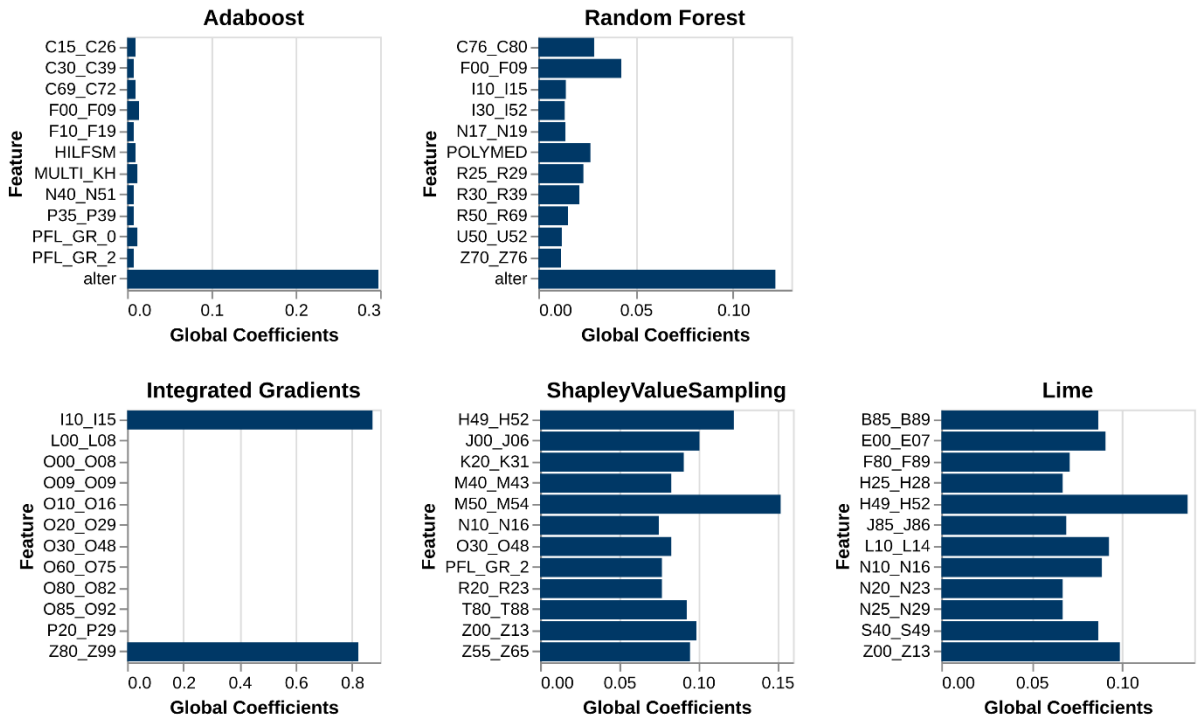

Abbildung 7-1. Balken Diagramme für Globale Feature Importance Werte der AdaBoost- und Random Forest-Modelle (oben), sowie Relevanzzuweisungen der XAI Methoden Integrated Gradients, LIME und ShapleyValueSampling (unten) für das Neuronale Netz Modell. Es werden immer nur die Feature mit den 12 größten Attributionen gezeigt.

In Abbildung 7-1 können wir sehen, dass für das AdaBoost- und Random Forest-Modell das Merkmal "Alter" den größten Einfluss auf die Vorhersage von Mortalität hat. Dies kann jedoch einfach darauf zurückgeführt werden, dass Alter die einzige metrische Variable ist und es den bereits beschriebenen Bias dafür gibt. Dies bedeutet, dass wir nur anhand dieser Werte für die Feature Importance nicht darauf schließen können, dass Alter die inhaltlich wichtigste Variable für die Vorhersage war. Alle anderen Variablen haben ansonsten relativ kleine Feature Importance Werte (<0.05). Für AdaBoost haben die Merkmale F00_09 (Organische, einschließlich symptomatischer psychischer Störungen), MULTI_KH (mehrere Krankenhausaufenthalte innerhalb von 6 Monaten), sowie PFL_GR_0 (kein Pflegegrad) die nächstgrößeren Feature Importance Werte. Für das Random Forest Modell haben die Merkmale F00_09 (s.o.), C76_C80 (Bösartige Neubildungen ungenau bezeichneter, sekundärer und nicht näher bezeichneter Lokalisationen) und POLYMED (Polymedikation) neben dem Alter die größten Feature Importance Werte.

Die Werte der Erklärbarkeitsmethoden für das Neuronale Netz unterschieden sich stark untereinander. Integrated Gradients hat nur positive Relevanzzuweisungen für I10_I15 (Hypertonie) und Z80_Z99 (Personen mit potenziellen Gesundheitsrisiken aufgrund der Familien- oder Eigenanamnese). Shapley Value Sampling hat die größte Relevanzzuweisung für M50_M54 (Sonstige Krankheiten der Wirbelsäule und des Rückens) gefolgt von H49_H52 (Affektionen der Augenmuskeln, Störungen der Blickbewegungen sowie Akkommodationsstörungen und Refraktionsfehler). H49_H52 hat die größte Relevanzzuweisung für die LIME-Methode gefolgt von Z00_Z13 (Personen, die das Gesundheitswesen zur Untersuchung und Abklärung in Anspruch nehmen).

7.3.2 Outcome Ungeplante Wiederaufnahme (Model 2)

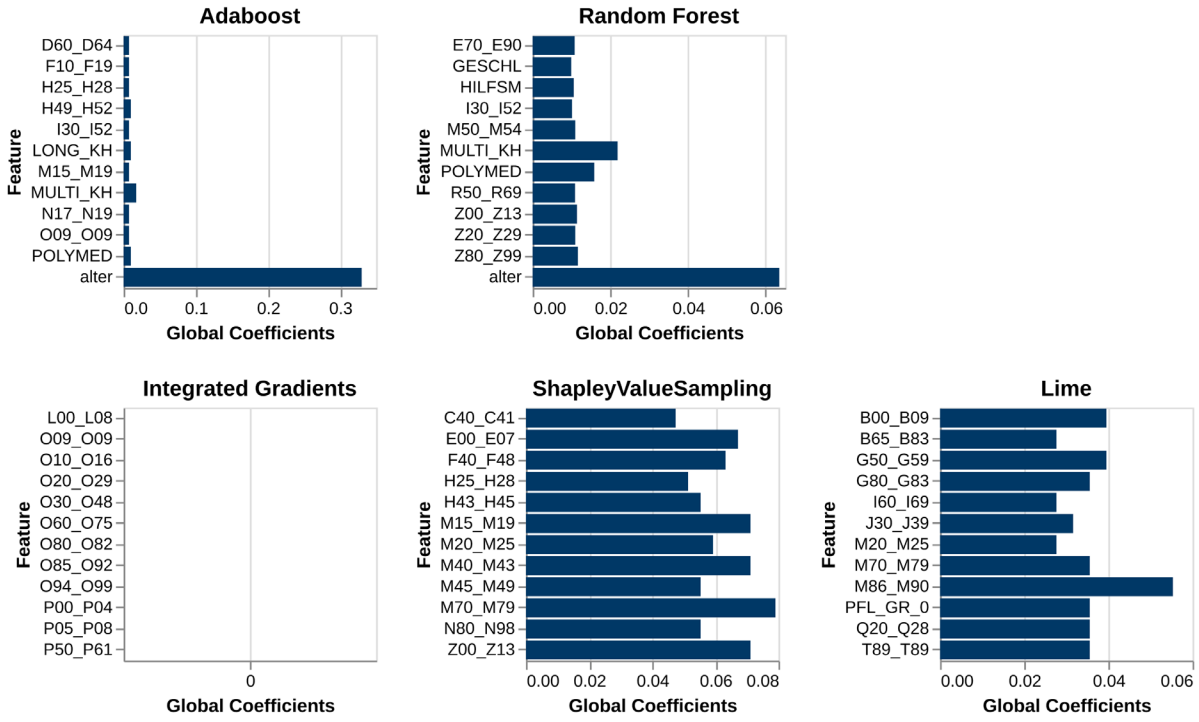

Abbildung 7-2. Balken Diagramme für Globale Feature Importance Werte der AdaBoost und Random Forest Modelle (oben), sowie Relevanzzuweisungen der XAI Methoden Integrated Gradients, LIME und ShapleyValueSampling (unten) für das Neuronale Netz Modell. Es werden immer nur die Feature mit den 12 größten Attributionen gezeigt.

Vergleichbar zur Abbildung 7-1 können wir auch in Abbildung 7-2 sehen, dass für das AdaBoost und Random Forest Modell Alter den größten Einfluss auf die Vorhersage von Ungeplanter Wiederaufnahme hat. Dies liegt jedoch nur daran, dass Alter die einzige metrische Variable ist und wir bloß von den Feature Importance Werten nicht darauf schließen können, dass Alter wirklich den größten (inhaltlichen) Einfluss auf die Vorhersage hatte. Für beide Modelle haben MULTI_KH (mehrere vorherige Krankenhausaufenthalte) und POLYMED (Polymedikation) die nächstgrößten Einflüsse auf die Vorhersage.

Integrated Gradients hat überhaupt keine positiven Relevanzzuweisungen für die Vorhersage von einer ungeplanten Wiederaufnahme. Shapley Value Sampling hat die größten Relevanzzuweisungen für M70_M79 (Sonstige Krankheiten des Weichteilgewebes), Z00_Z13 (Personen, die das Gesundheitswesen zur Untersuchung und Abklärung in Anspruch nehmen), M40_M43 (Deformitäten der Wirbelsäule und des Rückens) und M15_M19 (Arthrose). LIME hat die größten Relevanzzuweisungen für M86_M90 (Sonstige Osteopathien), B00_B09 (Virusinfektionen, die durch Haut- und Schleimhautläsionen gekennzeichnet sind) und G50_G59 (Krankheiten von Nerven, Nervenwurzeln und Nervenplexus).

7.4 Fazit

In diesem Kapitel haben wir gesehen, dass für Daten mit metrischen und kategoriellen Variablen die Feature Importance Methoden von AdaBoost-Modellen und Random Forest-Modellen wenig aussagekräftig sind. Metrische Variablen, wie im Anwendungsbeispiel das Merkmal Alter, oder Variablen mit vielen Kategorien werden algorithmisch bevorzugt. Trotzdem konnten wir erkennen, dass Variablen wie MULTI_KH (mehrfacher vorheriger Krankenhausaufenthalt) und POLYMED (Polymedikation) einen minimal größeren Einfluss auf die Vorhersagen als die Vorerkrankungsdiagnosen haben.

Von den Erklärbarkeitsmethoden hat Integrated Gradients am unzuverlässigsten funktioniert, mit nur wenigen bis gar keinen positiven Relevanzzuweisungen. Eine mögliche Erklärung für dieses Verhalten liegt an der Methode selbst, denn sie ist die einzige Methode, die den Gradienten des Modells sowie den Output des Models explizit zur Berechnung der Relevanzzuweisungen verwendet. Wenn diese sehr klein ausfallen, oder der Gradient sogar Null wird, dann werden auch die Relevanzzuweisungen sehr klein bzw. Null. Bei den anderen beiden Methoden, Shapley Value Sampling und LIME, kann dies nicht passieren, da diese auf Veränderungen in den Vorhersagen basieren, die durch leichte Änderungen der Testdaten hervorgerufen werden. Für diese Methoden gibt es einzelne Überlappungen in den Variablen mit den 12 größten Relevanzzuweisungen. Darunter finden sich fast ausschließlich Vorerkrankungsdiagnosen, die aufgrund ihrer Schwere erwartungsgemäß häufiger zu Wiederaufnahmen oder zum Tod führen. Beim Outcome Mortalität sind dies beispielsweise C-Diagnosen (Neubildungen), F-Diagnosen (Psychische und Verhaltensstörungen) und N-Diagnosen (Krankheiten des Urogenitalsystems). Beim Outcome Ungeplante Wiederaufnahme werden zudem M-Diagnosen (Krankheiten des Muskel-Skelett-Systems) und H-Diagnosen (Krankheiten des Auges) häufiger als relevante Merkmale detektiert.

Grundsätzlich gilt, dass die Feature Importances und Relevanzzuweisungen uns nur ein Indiz geben, welche Variablen für die Vorhersagen algorithmisch am wichtigsten waren. Zur fachlich-inhaltlichen Bewertung, ob zum Beispiel aus medizinischer Perspektive bestimmte Vorerkrankungen mit einem erhöhten Mortalitätsrisiko einhergehen, sind diese Verfahren allerdings nur sehr eingeschränkt zu empfehlen. Die Relevanzzuweisungen der Erklärbarkeitsmethoden könnten dies eher indizieren, da unseren Grafiken lokale Relevanzzuweisungen zugrunde liegen. Das heißt, dass wir für jeden Patienten im Testdatensatz genau sehen könnten, ob der spezifische Wert einer Variable positiven oder negativen Einfluss auf die Vorhersage von zum Beispiel Mortalität hätte. Es gilt jedoch zu beachten, dass sowohl den KI-Modellen als auch den Erklärbarkeitsmethoden stochastische Prozesse zugrunde liegen, wodurch die Relevanzzuweisungen für einen Patienten immer leicht unterschiedlich sein könnten und die Instabilität von Erklärbarkeitsmethoden generell ein bekanntes Problem ist. Letztendlich ist auch noch nicht geklärt, welche Erklärbarkeitsmethoden besser auf Routinedaten funktionieren als andere. Dafür sind große Studien mit Daten notwendig, bei denen bekannt ist, welche Variablen relevant sind. Ein großes Problem von erklärbarer KI im medizinischen Sektor ist, dass wir häufig nicht genau sagen können, welche Variablen von Bedeutung sind.

Die Ergebnisse der hier präsentierten Erklärbarkeitsmethoden sind ein erster Schritt, um einen Einblick in den Vorhersageprozess von KI-Modellen zu erlangen, aber zum jetzigen Zeitpunkt ermöglichen sie uns noch keine sicheren Erklärungen dafür, warum ein bestimmter Patient eine bestimmte Vorhersage bekommen hat.

Quellen

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, & Torsten Hothorn (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8:1±21. https://doi.org/10.1186/1471-2105-8-25

Erik Strumbelj and Igor Kononenko (2010). An efficient explanation of individual classifications using game theory. J. Mach. Learn. Res., 11:1±18.

Lloyd S. Shapley (1952). A Value for N-Person Games. *RAND Corporation*, Santa Monica, CA.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin (2016). ªwhy should I trust you?: Explaining the predictions of any classifier. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pages 97±101, San Diego, California. Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1602.04938

Mukund Sundararajan, Ankur Taly, and Qiqi Yan (2017). Axiomatic attribution for deep networks. https://doi.org/10.48550/arXiv.1703.01365

Ursula Neumann, Nikita Genze, and Dominik Heider (2017). EFS: an ensemble feature selection tool implemented as r-package and web-application. *BioData mining*, 10:1±9. https://doi.org/10.1186/s13040-017-0142-8

8 Fehlklassifikationen und Übertragbarkeit der Modelle

Wichtig für die Anwendbarkeit von Prognosemodellen ist, wie gut diese auf bestimmte Situationen übertragbar sind. Zur Überprüfung kann beispielweise die Performance der Modelle für unterschiedliche Zeiträume und Subgruppen untersucht und miteinander verglichen werden.
Im Anwendungsbeispiel zeigten sich bei einem Vergleich der besten Verfahren (log. Regression und AdaBoost) für einzelne Datenjahre und Subgruppen Schwankungen in der Prognosegüte (in Form des AUC-ROC-Werts), wobei es keinen wesentlichen Unterschied zwischen den beiden Verfahren gab.

Prognosemodelle liefern in der Regel keine perfekten Vorhersagen, sondern klassifizieren zumindest einen Teil der Fälle falsch ein (s. Kapitel 3 für die theoretischen Erläuterungen). Diese Fehlklassifikationen sind häufig unvorhersehbar (unvermeidbare Fehlklassifikationen), weil sie auf seltene Datenkonstellationen zurückzuführen sind (z. B. Patienten, bei denen ein Ereignis ohne dokumentierten Risikofaktor eintritt). Andererseits kann es aber auch sein, dass bestimmte Fälle systematisch häufiger fehlklassifiziert werden (kontrollierbare Fehlklassifikationen). Das ist beispielsweise der Fall, wenn bestimmte Subgruppen der Versicherten häufiger falsch klassifiziert werden. Auch bei der Übertragung eines trainierten Modells in die Praxis kann es zu einem Anstieg der Fehlklassifikationen kommen, insbesondere wenn sich die Trainingsdaten von den realen Daten unterscheiden. Typischerweise kann das allein dadurch zustanden kommen, dass die Trainingsdaten i.d.R. aus der Vergangenheit stammen (hier: Training der Modelle auf Routinedaten aus dem Jahr 2018), wohingegen die realen Daten, auf denen die Modelle angewendet werden sollen, später entstehen. Im vorliegenden Kapitel werden daher Ansätze beschrieben, um die Übertragbarkeit der Modelle auf zukünftige Daten oder bestimmte Subgruppen zu untersuchen.

Exemplarisch werden diese Ansätze in einem weiteren Abschnitt des Kapitels an den Modellen der logistischen Regression und AdaBoost, welches sich in den vorangestellten Kapiteln als bestes ML-Verfahren erwiesen hat, demonstriert. Ziel dabei ist es, zu untersuchen, ob sich die prognostische Güte der beiden Modelle in Bezug auf einzelne Versicherte bzw. Versichertenpopulationen bedeutend voneinander unterscheidet.

8.1 Einleitung und theoretische Überlegungen

8.1.1 Fehlklassifikationen und Übertragbarkeit der Modelle

Prognosemodelle werden typischerweise an bestehenden bzw. an in der Vergangenheit entstandenen Daten trainiert, sollen aber genutzt werden, um zukünftige Ereignisse anhand von aktuellen Daten vorherzusagen. Mitunter kann dies zu Problemen führen, wenn sich die neuen Daten systematisch von den historischen Daten unterscheiden. In solchen Fällen machen die Modelle häufig falsche Vorhersagen, weil sie nicht in der Lage sind, auf neue Situationen zu generalisieren. Im Kontext von KI-THRUST haben wir Prognosemodelle erstellt, die Mortalität und Ungeplante Wiederaufnahmen nach Krankenhausentlassung vorhersagen und auf Daten aus dem Jahr 2018 trainiert und getestet wurden. Beispielsweise liegt die Vermutung nah, dass die Corona-Pandemie, die vor allem im Jahr 2020 mit gravierenden Auswirkungen auf die Gesundheitsversorgung der Versicherten einherging, im Rahmen der Modellierung dazu führt, dass sich die Fälle aus dem Jahr 2020 deutlich schlechter mit den trainierten Modellen aus dem Jahr 2018 vorhersagen lassen als beispielsweise die Fälle aus dem letzten Vor-Corona-Jahr 2019. Um diese Hypothese zu testen, wurden die finalen Prognosemodelle genutzt, um Daten aus dem Jahr 2019 und dem Jahr 2020 vorherzusagen und mögliche Unterschiede in der Prognosegüte zu untersuchen.

Darüber hinaus kann es sein, dass Fälle mit bestimmten Charakteristiken grundsätzlich häufiger fehlklassifiziert werden als alle anderen Fälle. Beispielsweise könnten bei der Prognose der Mortalität bestimmte Altersgruppen (z. B. besonders alte oder junge Menschen) häufiger fehlklassifiziert werden als Menschen mittleren Alters. Daher wurde untersucht, ob sich etwaige, systematisch erhöhte Fehlklassifikationsraten bei den Prognosemodellen aus KI-THRUST für bestimmte Subgruppen feststellen lassen. Dazu wurden hypothesengeleitet Subgruppen gebildet, die einen Einfluss auf die Fehlklassifikationsrate haben könnten. Neben Subgruppen auf Basis von Alter, Geschlecht und Charakteristika des Krankenhausfalls (Notfall vs. Normalfall) wurden auch Subgruppen auf Basis der potenziellen Datenverfügbarkeit gebildet. Der Grundgedanke bei der potenziellen Datenverfügbarkeit ist, dass möglicherweise Fälle von Personen, von denen nur wenig Daten aus dem Gesundheitssystem vorliegen, schlechter vorherzusagen sind. Der Grund dafür, dass für einzelne Personen nur wenig Daten vorliegen, kann sein, dass diese Personen seltener krank sind und daher das Gesundheitssystem auch seltener in Anspruch nehmen müssen. Es kann aber auch sein, dass diese Personen das Gesundheitssystem nicht in Anspruch nehmen wollen oder möglicherweise anderweitig behandelt werden. In beiden Fällen sollten für diese Person wenig bis keine Eintragungen im ambulant-ärztlichen Bereich vorliegen (in dem typischerweise bei den meisten Personen zumindest innerhalb eines Jahres Daten vorliegen). Eine mögliche Gruppe, für die dies gelten kann, sind Person in stationären Pflegeeinrichtungen. Da die Angaben zur Inanspruchnahme stationärer Pflegeleistungen in den KI-THRUST-Daten enthalten waren, konnte diese Versichertengruppe abgebildet und gezielt untersucht werden.

8.1.1 Vergleich der Prognosen zwischen logistischer Regression und AdaBoost

Neben der generellen Übertragbarkeit der Modelle auf zukünftige Jahre und auf bestimmte Subgruppen, war für uns auch ein direkter Vergleich der logistischen Regressionsmodelle mit den ML-Verfahren auf Personenebene interessant. Ziel war es, zu untersuchen, ob die jeweils besten Modelle für einzelne Versicherte zu ähnlichen Einschätzungen gelangen, d. h. ob eine Person, die bei der logistischen Regression eine hohe Wahrscheinlich für das Eintreten des Outcomes hat, dies auch bei den ML-Verfahren aufweist. Die Ergebnisse der Modelltestung (s. Kapitel 7.3) zeigen, dass die jeweils besten Modelle (Logistische Regression und AdaBoost, jeweils Modell 2) eine sehr ähnliche Gesamtperformance aufweisen, was darauf hindeutet, dass die Vorhersagen auf Basis sehr ähnlicher Parameter stattfinden. Wenn die Verfahren ihre Vorhersagen auf Basis ähnlicher Parameter treffen, sollten sie auch für einzelne Versicherte zu relativ gleichen Ergebnisse gelangen und die jeweiligen Reihenfolgen

der Versicherten (Reihung von niedrigste bis höchste Eintrittswahrscheinlichkeit für ein Outcome) sollten sich annähern.

8.2 Methoden

8.2.1 Übertragbarkeit auf zukünftige Datenjahre

Um die Übertragbarkeit der auf dem Datenjahr 2018 trainierten Modelle auf nachfolgende bzw. „zukünftige" Jahre zu untersuchen, wurden das beste logistische Regressionsmodell sowie das beste ML-Modell (s. Kapitel 7.3) genutzt, um die Prognosegüte für die beiden Outcomes mit Daten aus dem Jahr 2019 und 2020 zu testen und zu vergleichen. Für die logistische Regression wurde das Datenmodell 2 (M2) genutzt, als ML-Verfahren wurde AdaBoost – ebenfalls mit dem Datenmodell 2 (M2) – für Vergleichszwecke herangezogen. Die Daten aus den Jahren 2019 und insbesondere 2020 unterscheiden sich in manchen Aspekten signifikant von dem Jahr 2018 (s. Kapitel 6.2). Sowohl im Jahr 2019 als auch im Jahr 2020 gab es mehr Versicherte mit Pflegegrad und mehr Hilfsmittelverordnungen. Im Jahr 2020 waren die Versicherten darüber hinaus signifikant älter und das Outcome Mortalität trat häufiger auf. Als Maß für die Prognosegüte wurden die Receiver Operating Characteristic (ROC-Kurve) und die Precision-Recall-Kurve für alle Datenjahre erstellt und die dazugehörigen Werte für die Fläche unter der Kurve (AUC-ROC und AUC-PR) miteinander verglichen.

8.2.2 Anwendbarkeit in Subgruppen

Um die Prognosegüte der Modelle in unterschiedlichen Subgruppen zu untersuchen, wurden das beste logistische Regressionsmodell sowie das AdaBoost-Verfahren als bestes ML-Modell (jeweils Datenmodell M2) genutzt, um für jede Subgruppe die Receiver Operating Characteristic (ROC-Kurve) zu berechnen und den Wert für die Fläche unter der Kurve (AUC-ROC) mit Angabe des 95 %-Konfidenzintervalls miteinander zu vergleichen. Für die Variablen Geschlecht (männlich/weiblich) und Alter (in 20-Jahres-Altersgruppen) wurden Subgruppen gebildet, sowie für die Art des Krankenhausaufenthalts des Indexfalls (Normalfall vs. Notfall). Um die Auswirkung der Datenverfügbarkeit zu untersuchen, wurden Patienten/Patientinnen aus Pflegeheimen untersucht. Dazu wurden Patienten/Patientinnen, die im Jahr vor Aufnahme ins Krankenhaus mindestens 90 Tage in stationärer Pflege waren, verglichen mit Patienten/Patientinnen, die gar nicht oder kürzer in stationärer Pflege waren.

8.2.3 Vergleich der Prognosen zwischen logistischer Regression und AdaBoost

Für jeden Versicherten geben die Verfahren (Logistische Regression und ML-Verfahren) einen Score-Wert aus. Bei der logistischen Regression entspricht dieser der geschätzten Wahrscheinlichkeit für das Eintreten des jeweiligen Outcomes. Für AdaBoost (als im vorliegenden Beispiel besten ML-Verfahren) wird durch die Gewichtung beim Modelltraining kein Score-Wert erstellt, der im gleichen Maße im Sinne einer Eintrittswahrscheinlichkeit interpretiert werden kann. Die Score-Werte sind somit nicht direkt miteinander vergleichbar. Dennoch lässt sich annehmen, dass – bei übereinstimmenden Prognosen der beiden Verfahren – Personen mit einer hohen geschätzten Eintrittswahrscheinlichkeit bei der logistischen Regression genauso auch einen vergleichsweise hohen Score-Wert bei AdaBoost aufweisen und Personen mit einer niedrigen geschätzten Eintrittswahrscheinlichkeit bei der logistischen Regression genauso auch einen vergleichsweise niedrigen Score-Wert bei AdaBoost aufweisen. Foglich sollte zumindest die Rangreihenfolge der Score-Werte vergleichbar sein. Um dies zu untersuchen, haben wir die Score-Werte für beide Verfahren jeweils nach Rängen geordnet (wobei hohe Ränge eine vergleichsweise hohe Eintrittswahrscheinlichkeit für das Outcome bedeuten) und diese gegeneinander aufgetragen. Als Vergleichsmodell wurde dabei jeweils wieder für beide Outcomes das Datenmodell M2 gewählt.

8.3 Ergebnisse zur Fehlklassifikation und Übertragbarkeit

8.3.1 Prognosegüte in zukünftigen Jahren

Um die Übertragbarkeit der Modelle auf die Jahre 2019 und 2020 zu untersuchen, wurden die auf den Daten aus dem Jahr 2018 entwickelten Modelle auf Daten aus dem Jahr 2019 und 2020 getestet. Die Prognosegüte wurde anhand der Fläche unter der ROC-Kurve (AUC-ROC) verglichen. Detaillierte Ergebnisse sind in Abbildung 8-1 sowie in Tabelle 8-1 zu finden.

Für das Outcome Mortalität zeigten sich für das Jahr 2020 etwas geringere AUC-ROC-Werte im Vergleich zu den Jahren 2018 und 2019, sowohl bei der logistischen Regression als auch bei dem Maschinellen Lernverfahren (AdaBoost; s. auch Abbildung 8-1). Bei dem Outcome Ungeplante Wiederaufnahmen zeigten sich keine wesentlichen Unterschiede zwischen den Datenjahren.

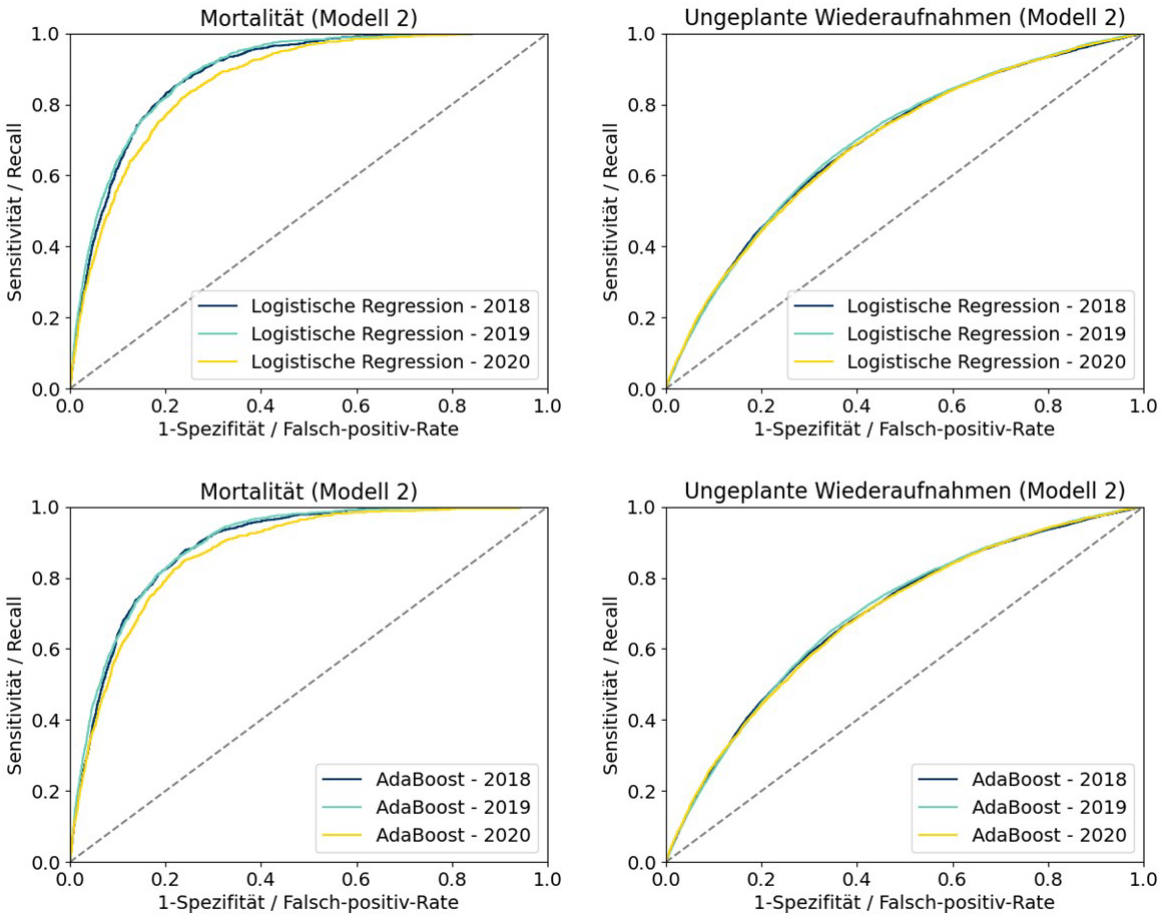

Abbildung 8-1. Receiver-Operating Characteristics (ROC-Kurven) für die Testdaten der Jahr 2018, 2019 und 2020 im Vergleich

Tabelle 8-1. Prognosegüte in zukünftigen Jahren: Fläche unter der Receiver-Operating Characteristic (AUC-ROC) und 95 %-Konfidenzintervall (KI), sowie Anzahl Versicherte (N) und Prävalenz des Outcomes je Datenjahr

Verfahren	Datenjahr	AUC-ROC [95 %-KI]	N	Prävalenz Outcome
Outcome: Mortalität (Modell 2)				
Logistische Regression	2018	0,888 [0,880; 0,895]	118.767	0,98 %
	2019	0,893 [0,885; 0,900]	118.640	0,97 %
	2020	0,866 [0,856; 0,876]	76.850	1,15 %
AdaBoost	2018	0,889 [0,882; 0,896]	118.767	0,98 %
	2019	0,892 [0,885; 0,900]	118.640	0,97 %
	2020	0,871 [0,861; 0,881]	76.850	1,15 %
Outcome: Ungeplante Wiederaufnahmen (Modell 2)				
Logistische Regression	2018	0,693 [0,688; 0,698]	118.767	8,18 %
	2019	0,697 [0,691; 0,702]	118.640	8,30 %
	2020	0,692 [0,685; 0,698]	76.850	7,93 %
AdaBoost	2018	0,694 [0,689; 0,699]	118.767	8,18 %
	2019	0,697 [0,692; 0,703]	118.640	8,30 %
	2020	0,692 [0,686; 0,699]	76.850	7,93 %

8.3.2 Prognosegüte in Subgruppen

Um die Güte der Prognosemodelle in den Subgruppen zu vergleichen, wurde die Fläche unter der ROC-Kurve (AUC-ROC) je Subgruppe berechnet. Detaillierte Ergebnisse sind in Tabelle 8-2 zu finden.

Tabelle 8-2. Subgruppenanalysen: Fläche unter der Receiver-Operating Characteristic (AUC-ROC) und 95 %-Konfidenzintervall (KI), sowie Anzahl Versicherte (N) und Prävalenz des Outcomes je Subgruppe

Subgruppe	Ausprägung	Verfahren	AUC-ROC [95 %-KI]	N	Prävalenz Outcome
Outcome: Mortalität (Modell 2)					
Geschlecht	Weiblich	Logistische Regression	0,900 [0,890; 0,910]	59.968	0,98 %
	Männlich		0,874 [0,862; 0,886]	58.799	0,97 %
	Weiblich	AdaBoost	0,902 [0,892; 0,911]	59.968	0,98 %
	Männlich		0,875 [0,863; 0,886]	58.799	0,97 %
Altersgruppe	0 - 20 Jahre*	Logistische Regression	*0,567 [0,041; 1,000]*	*8.474*	*0,02 %*
	21 - 40 Jahre*		*0,997 [0,996; 0,998]*	*14.854*	*0,01 %*
	41 - 60 Jahre		0,904 [0,864; 0,944]	25.043	0,24 %
	61 - 80 Jahre		0,833 [0,812; 0,854]	45.137	0,76 %
	> 80 Jahre		0,764 [0,749; 0,780]	25.259	2,98 %
	0 - 20 Jahre*	AdaBoost	*0,943 [0,843; 1,000]*	*8.474*	*0,02 %*
	21 - 40 Jahre*		*0,998 [0,997; 0,998]*	*14.854*	*0,01 %*

Subgruppe	Ausprägung	Verfahren	AUC-ROC [95 %-KI]	N	Prävalenz Outcome
	41 - 60 Jahre		0,915 [0,883; 0,947]	25.043	0,24 %
	61 - 80 Jahre		0,837 [0,817; 0,858]	45.137	0,76 %
	> 80 Jahre		0,763 [0,747; 0,778]	25.259	2,98 %
Krankenhausauf-enthalt	Normallfall	Logistische Regression	0,897 [0,880; 0,914]	59.168	0,46 %
	Notfall		0,866 [0,856; 0,875]	58.570	1.50 %
	Normallfall	AdaBoost	0,900 [0,885; 0,916]	59.168	0,46 %
	Notfall		0,866 [0,857; 0,875]	58.570	1,50 %
Pflegeheim	Ja (> 90 Tagen)	Logistische Regression	0,729 [0,695; 0,763]	3.398	5,12 %
	Nein (≤ 90 Tage)		0,886 [0,878; 0,895]	115.369	0,85 %
	Ja (> 90 Tagen)	AdaBoost	0,720 [0,686; 0,754]	3.398	5,12 %
	Nein (≤ 90 Tage)		0,888 [0,880; 0,896]	115.369	0,85 %
Outcome: Ungeplante Wiederaufnahmen (Modell 2)					
Geschlecht	Weiblich	Logistische Regression	0,688 [0,680; 0,695]	59.968	7,80 %
	Männlich		0,697 [0,690; 0,705]	58.799	8,57 %
	Weiblich	AdaBoost	0,689 [0,681; 0,696]	59.968	7,80 %
	Männlich		0,698 [0,691; 0,706]	58.799	8,57 %
Altersgruppe	0 - 20 Jahre	Logistische Regression	0,698 [0,669; 0,728]	8.474	4,84 %
	21 - 40 Jahre		0,674 [0,653; 0,695]	14.854	5,12 %
	41 - 60 Jahre		0,691 [0,676; 0,707]	25.043	5,49 %
	61 - 80 Jahre		0,662 [0,653; 0,670]	45.137	8,99 %
	> 80 Jahre		0,612 [0,602; 0,623]	25.259	12,33 %
	0 - 20 Jahre	AdaBoost	0,711 [0,683; 0,740]	8.474	4,84 %
	21 - 40 Jahre		0,679 [0,659; 0,700]	14.854	5,12 %
	41 - 60 Jahre		0,695 [0,679; 0,711]	25.043	5,49 %
	61 - 80 Jahre		0,662 [0,653; 0,670]	45.137	8,99 %
	> 80 Jahre		0,613 [0,603; 0,624]	25.259	12,33 %
Krankenhausauf-enthalt	Normallfall	Logistische Regression	0,687 [0,677; 0,696]	59.168	5,49 %
	Notfall		0,679 [0,673; 0,686]	58.570	10,93 %
	Normallfall	AdaBoost	0,689 [0,680:0; 699]	59.168	5,49 %
	Notfall		0,679 [0,672; 0,686]	58.570	10,93 %
Pflegeheim	Ja (> 90 Tagen)	Logistische Regression	0,553 [0,526; 0,581]	3.398	14,30 %
	Nein (≤ 90 Tage)		0,694 [0,688; 0,699]	115.369	8,00 %
	Ja (> 90 Tagen)	AdaBoost	0,555 [0,527; 0,583]	3.398	14,30 %
	Nein (≤ 90 Tage)		0,695 [0,689; 0,700]	115.369	8,00 %

*Ergebnisse aufgrund geringer Betroffenenzahlen nicht inhaltlich interpretierbar

Geschlecht

Für Subgruppen unterteilt nach Geschlecht zeigten sich für das Outcome Mortalität höhere AUC-ROC-Werte für Frauen im Vergleich zu Männern. Für das Outcome Ungeplante Wiederaufnahmen hingegen waren die AUC-ROC-Werte für Männer etwas höher als die für Frauen, allerdings mit deutlichen kleineren Unterschieden zwischen den Gruppen. Die Ergebnismuster waren für die logistische Regression und das Maschinelle Lernverfahren (AdaBoost) gleich, s. Abbildung 8-2.

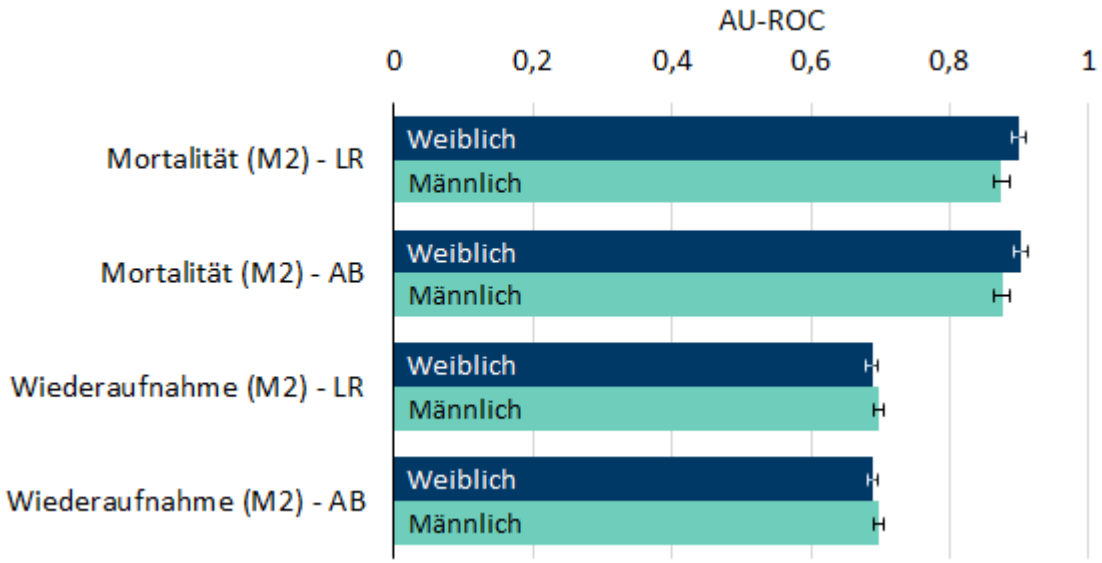

Abbildung 8-2. Subgruppenanalyse für die Variable Geschlecht: Fläche unter der Receiver-Operating Characteristic (AUC-ROC) und 95 %-Konfidenzintervall je Subgruppe für die logistische Regression (LR) und das ML-Verfahren AdaBoost (AB)

Altersgruppen

In Altersgruppen zeigten sich insbesondere im Hinblick auf das Outcome Mortalität deutliche Unterschiede mit niedrigeren AUC-ROC-Werten in den höheren Altersgruppen (s. Abbildung 8-3). Dabei sind Ergebnisse zum Outcome Mortalität in den beiden Altersgruppen bis 40 Jahre aufgrund einer geringen Betroffenenzahl (mit 1 bis 2 Todesfällen je Gruppe) inhaltlich nicht interpretierbar und sollten hier nicht beachtet werden, zumal die verwendeten Teststatistiken hier offensichtlich auch zur Ermittlung inadäquater Vertrauensbereiche führen können.

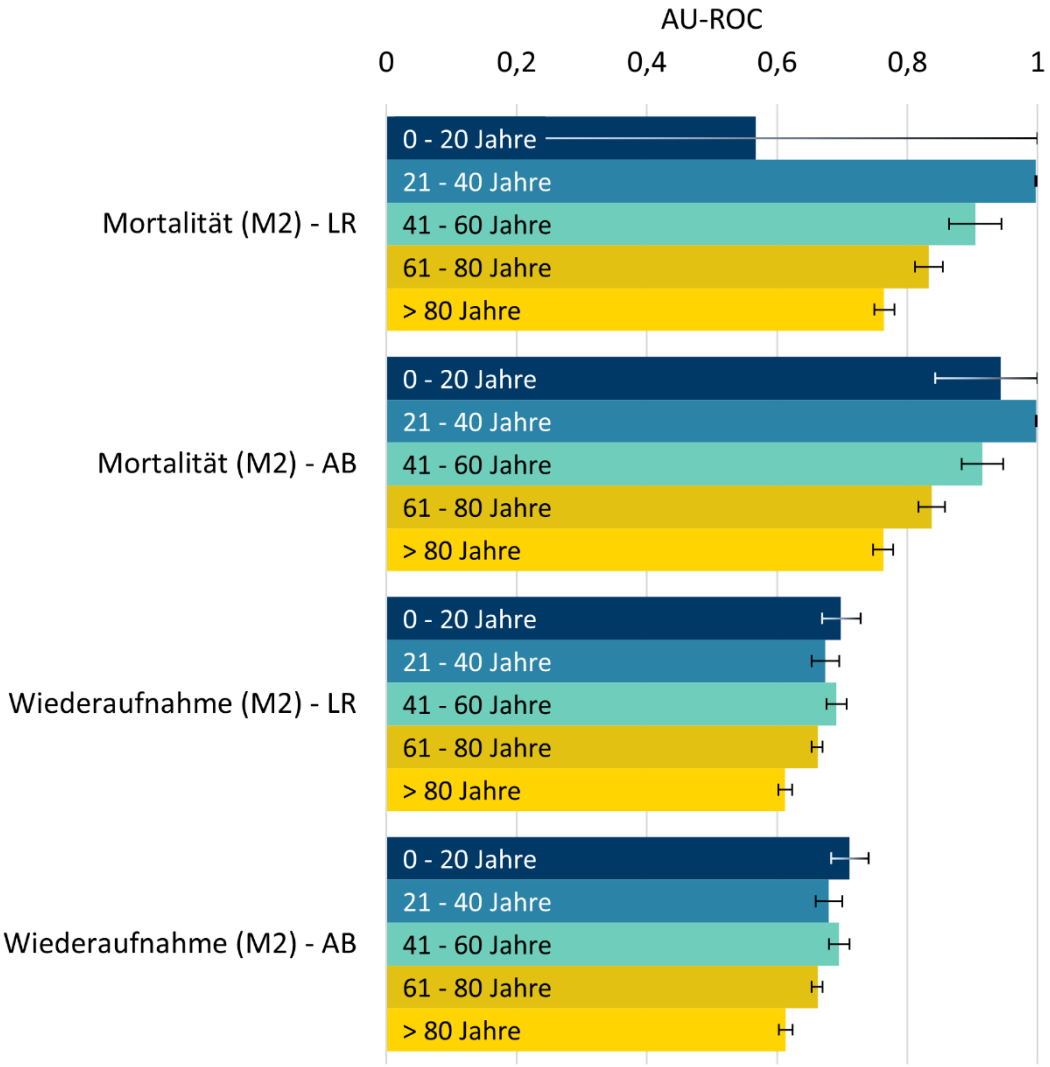

Abbildung 8-3. Subgruppenanalyse für die Variable Alter: Fläche unter der Receiver-Operating Characteristic (AUC-ROC) und 95 %-Konfidenzintervall je Subgruppe für die logistische Regression (LR) und das ML-Verfahren AdaBoost (AB)

Krankenhausfall: Normalfall vs. Notfall

Für Subgruppen unterteilt nach der Art des Krankenhausaufenthaltes (Normaler Krankenhausaufenthalt oder notfallmäßige Aufnahme) zeigten sich insbesondere für das Outcome Mortalität höhere AUC-ROC-Werte für normale Krankenhausaufenthalte im Vergleich zu Notfallaufnahmen. Diese Tendenz zeigte sich auch für das Outcome Ungeplante Wiederaufnahmen, allerdings waren die Unterschiede hier deutlich kleiner und im zu vernachlässigenden Bereich. Die Ergebnismuster waren für die logistische Regression und das Maschinelle Lernverfahren (Ada Boost) gleich (s. Abbildung 8-4).

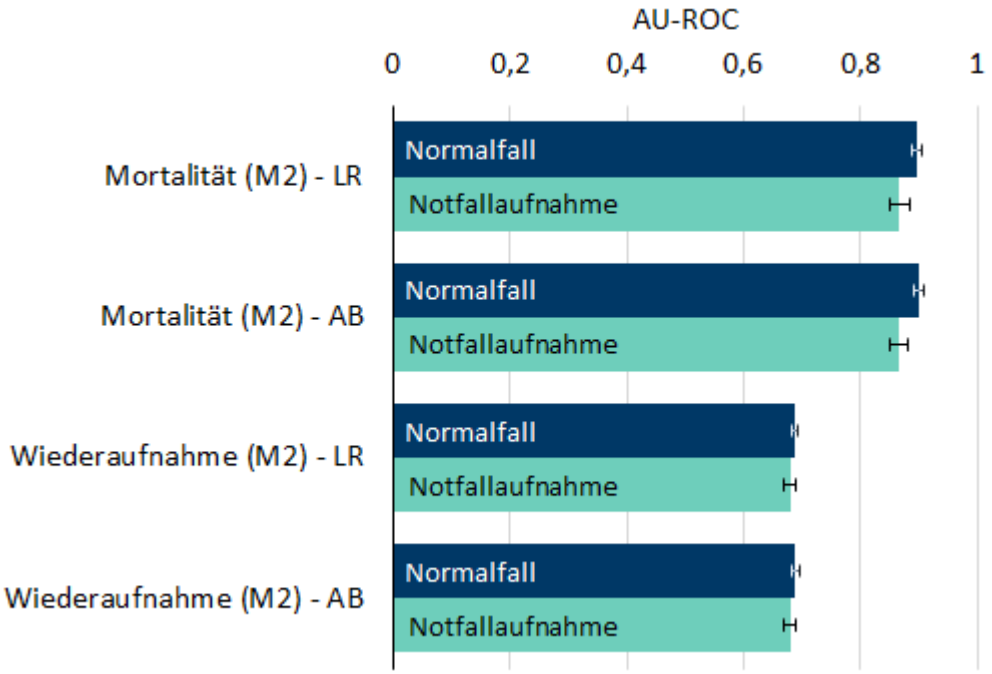

Abbildung 8-4. Subgruppenanalyse für die Variable Art des Krankenhausaufenthalts: Fläche unter der Receiver-Operating Characteristic (AUC-ROC) und 95 %-Konfidenzintervall je Subgruppe für die logistische Regression (LR) und das ML-Verfahren AdaBoost (AB)

Pflegeheim

Für Subgruppen unterteilt nach Pflegeheimaufenthalt zeigten sich deutliche Unterschiede in den ROC-AUC-Werten für beide Outcomes: Die Werte waren deutlich geringer für Personen, die im Jahr vor dem Krankenhausaufenthalt 90 Tage oder länger in einem Pflegeheim waren im Vergleich zu Personen, die kürzer oder gar nicht in einem Pflegeheim waren. Die Ergebnismuster waren für die logistische Regression und das Maschinelle Lernverfahren (Ada Boost) gleich (s. Abbildung 8-5).

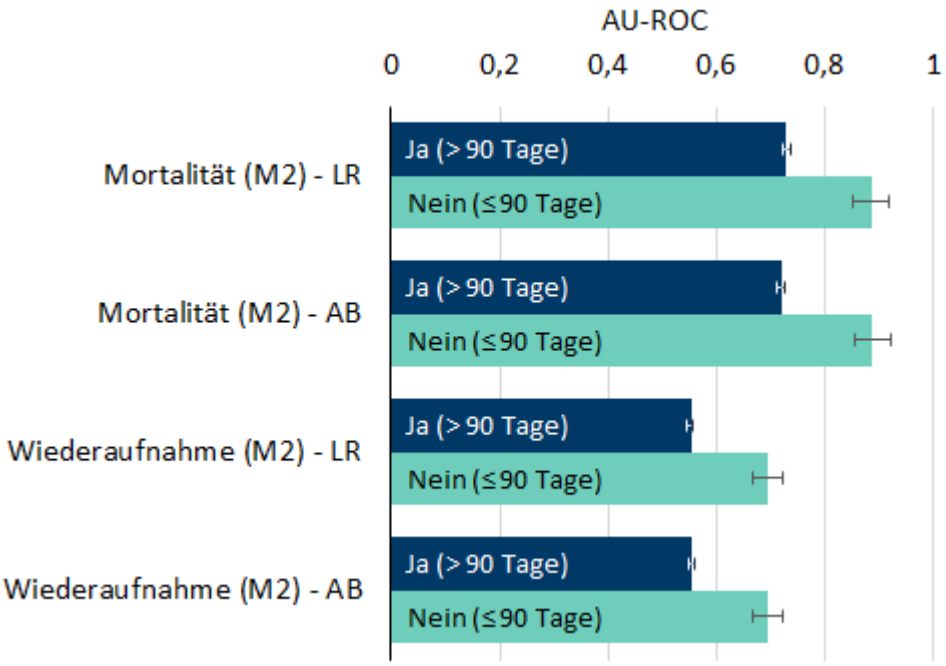

Abbildung 8-5. Subgruppenanalyse für die Variable Pflegeheim: Fläche unter der Receiver-Operating Characteristic (AUC-ROC) und 95 %-Konfidenzintervall je Subgruppe für die logistische Regression (LR) und das ML-Verfahren AdaBoost (AB)

8.3.3 Vergleich der Prognosen zwischen logistischer Regression und AdaBoost

Ungeplante Wiederaufnahmen

Beim Auftragen der Score-Ränge für logistische Regression und AdaBoost in einem Streudiagramm zeigt sich, dass die Punkte um die Gerade streuen, die zu erwarten wäre, wenn alle Versicherten bei beiden Verfahren den gleichen Rang hätten. (s. Abbildung 8-6). Die Modelle sagen also nicht exakt die gleiche Reihung bei den Versicherten voraus. Die Streuung scheint dabei bei den unteren Rängen, sprich bei den Fällen mit vergleichsweise geringen Eintrittswahrscheinlichkeiten, größer zu sein als bei den oberen Rängen mit höheren Eintrittswahrscheinlichkeiten. Allerdings scheint es keine erkennbaren Cluster an Punkten zu geben, die einer der beiden Klassifikationsgruppen zugeordnet werden können und die deutlich von der Erwartung abweichen. Daher ist nicht davon auszugehen, dass eine systematische Fehlklassifikation vorliegt, die tiefergehend untersucht werden müsste.

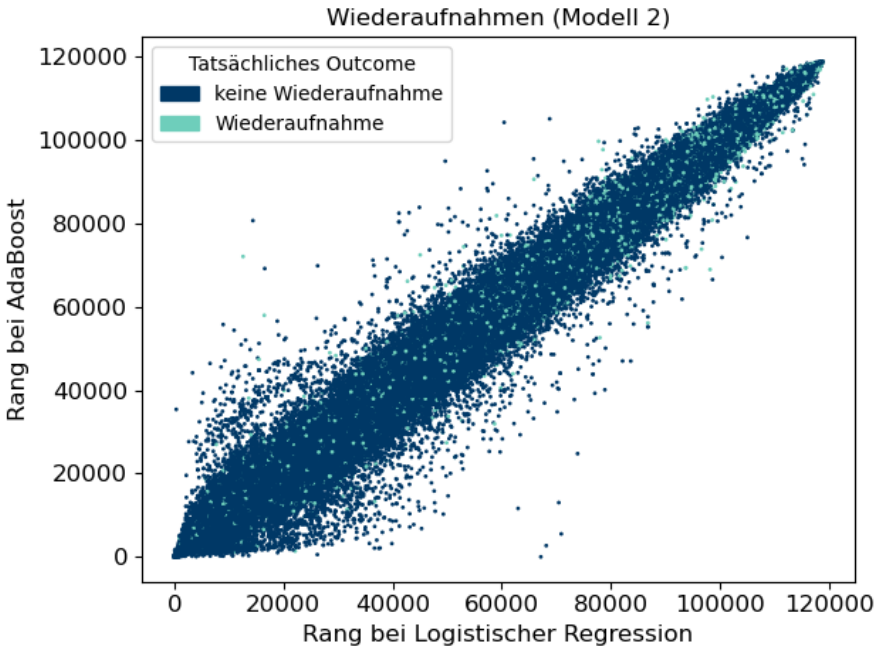

Abbildung 8-6. Scatterplot für die nach Rängen sortierten Score-Werte der Modelle M2 für die logistische Regression und AdaBoost für das Outcome Ungeplante Wiederaufnahmen (wobei hohe Ränge eine vergleichsweise hohe Eintrittswahrscheinlichkeit für das Outcome bedeuten).

Mortalität

Beim Auftragen der Score-Ränge für logistische Regression und AdaBoost zeigt sich, dass die Punkte um die Gerade streuen, die zu erwarten wäre, wenn alle Versicherten bei beiden Verfahren den gleichen Rang hätten. Allerdings gibt es auch eine Gruppe an Punkten, die deutlich von dieser Geraden abweicht, nämlich Versicherte, die bei AdaBoost einen sehr niedrigen Rang (< 6.000) aufweisen, bei der logistischen Regression aber einen höheren Rang von über 20.000 (s. Abbildung 8-7, A-links). Diese Subgruppe umfasst 1.100 Personen, von denen keine Person tatsächlich verstirbt (s. Abbildung 8-7, A-rechts und B). AdaBoost scheint hier also mit der Vorhersage niedriger Ränge näher an der Realität zu liegen als die logistische Regression.

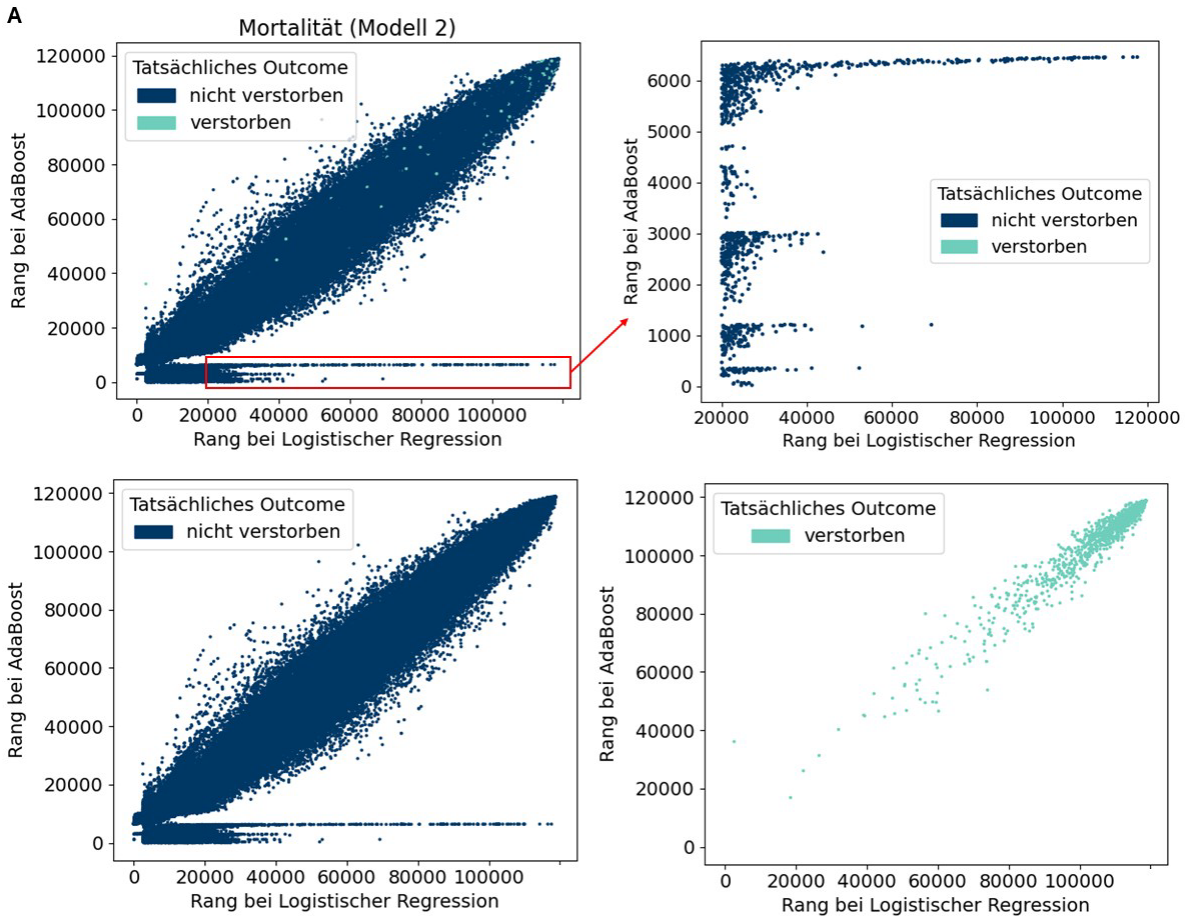

Abbildung 8-7. Scatterplot für die nach Rängen sortierten Score-Werte der Modelle M2 für die logistische Regression und AdaBoost für das Outcome Mortalität (wobei hohe Ränge eine vergleichsweise hohe Eintrittswahrscheinlichkeit für das Outcome bedeuten). A: Für die gesamte Stichprobe der Testdaten 2018 (links) und für einen Ausschnitt bestimmter Ränge (rechts). B: Aufgeteilt nach dem tatsächlichen Outcome, nicht verstorben (links) und verstorben (rechts).

Betrachtet man die Personen dieser Subgruppe näher, so fällt auf, dass der Frauenanteil deutlich erhöht ist und das Durchschnittsalter deutlich verringert in der Subgruppe im Vergleich zur Gesamtstichprobe (s. Tabelle 8-3). Die Personen dieser Gruppe sind also im Vergleich besonders jung und weiblich. In Anbetracht der häufigsten Diagnosen fällt zudem auf, dass in dieser Subgruppe schwangerschaftsassoziierte Diagnosen deutlich häufiger sind als in der Gesamtstichprobe (s. Tabelle 8-3).

Bei genauer Betrachtung des logistischen Regressionsmodells wird ersichtlich, dass schwangerschaftsassoziierte Diagnosen – anders als bei AdaBoost-Modellierungen – nicht im Modell enthalten sind. Durch die schrittweise Vorwärtsselektion (stepwise-forward-selection) wurden die Diagnosen hier (aufgrund mangelnder Signifikanz) nicht berücksichtigt. Dies scheint die Erklärung für die Abweichung zwischen dem logistischen Regressionsmodell und AdaBoost zu sein und dazu zu führen, dass AdaBoost für diese Subgruppe eine adäquatere Einschätzung trifft.

Tabelle 8-3. Häufigkeit bzw. Mittelwert ausgewählter Prädiktoren in der im Scatterplot identifizierten Subgruppe verglichen mit den Gesamttestdaten 2018

Merkmal	Subgruppe	Gesamttestdaten
Alter in Jahren (Mittelwert; Wertebereich)	39,7 (0-98)	61,2 (0-106)
Geschlecht (Anteil weiblich)	84,8 %	50,5 %
Pflegegrad (Anteil mit Pflegegrad)	5,5 %	14,4 %
Polymedikation (Anteil)	15,0 %	38,0 %
Multiple KH-Aufenthalte (Anteil)	15,0 %	16,2 %
Lange KH-Aufenthalte (Anteil)	7,6 %	8,9 %
Auswahl der 10 in der Subgruppe am häufigsten vertretenen Diagnosen:		
Z30-Z39: Personen, die das Gesundheitswesen im Zusammenhang mit Problemen der Reproduktion in Anspruch nehmen	18,1 %	0,9 %
Z80-Z90: Personen mit potentiellen Gesundheitsrisiken aufgrund der Familien- oder Eigenanamnese und bestimmte Zustände, die den Gesundheitszustand beeinflussen	16,1 %	28,7 %
O20-O29: Sonstige Krankheiten der Mutter, die vorwiegend mit der Schwangerschaft verbunden sind	12,9 %	0,8 %
O30-O39: Betreuung der Mutter im Hinblick auf den Fetus und die Amnionhöhle sowie mögliche Entbindungskomplikationen	12,5 %	0,6 %
Z00-Z13: Personen, die das Gesundheitswesen zur Untersuchung und Abklärung in Anspruch nehmen	11,8 %	11,0 %
I10-I15: Hypertonie	9,9 %	30,5 %
O94-O99: Sonstige Krankheitszustände während der Gestationsperiode, die anderenorts nicht klassifiziert sind	9,5 %	0,5 %
R50-R69: Allgemeinsymptome	9,5 %	12,7 %
Z70-Z76: Personen, die das Gesundheitswesen aus sonstigen Gründen in Anspruch nehmen	7,9 %	9,4 %
R10-R19: Symptome, die das Verdauungssystem und das Abdomen betreffen	7,8 %	10,9 %

8.4 Fazit

In diesem Kapitel wurde die Übertragbarkeit der berechneten Modelle sowie ihre Anwendbarkeit in ausgewählten Subgruppen untersucht. Dazu wurde die Performance der Modelle für unterschiedliche Zeiträume und Versichertengruppen analysiert und das logistische Regressionsmodell mit dem ML-Modell des AdaBoost-Verfahrens verglichen.

Bei der Vorhersage von Daten aus zukünftigen Jahren zeigten sich schlechtere Vorhersagen nur für das Outcome Mortalität im Jahr 2020 – die Vorhersagegüte war hier schlechter als in den Jahren 2018 und 2019. Dies könnte auf die Corona-Pandemie zurückzuführen sein, wobei vermutlich weniger der unmittelbare Effekt der Covid-19-Infektion auf die Mortalität eine Rolle gespielt haben dürfte, sondern vielmehr die stark veränderte Patientenstruktur in den Krankenhäusern – bedingt durch das Aufschieben elektiver Eingriffe und den Fokus auf die Versorgung akuter (Not-)Fälle. Für das Outcome Ungeplante Wiederaufnahmen zeigte sich hingegen kein Unterschied.

Bei der Analyse der Subgruppen ließen sich bei den Modellen Unterschiede in der Performance feststellen. Insbesondere innerhalb von höheren Altersgruppen und bei Pflegeheimbewohnern war die Prognosegüte (in Form des AUC-ROC-Werts) sowohl bei der logistischen Regression als auch bei AdaBoost schlechter. Dies könnte darauf zurückzuführen sein, dass anderweitige Risikofaktoren innerhalb von Gruppen mit ohnehin deutlich erhöhten Risiken nur noch eingeschränkt zur weiteren Differenzierung von Risiken beitragen. Wird beispielsweise eine in hohem Alter eher gewöhnliche Diagnose bereits bei jungen Menschen gestellt, dürfte sie das Risiko maßgeblich verändern, wohingegen sie bei älteren Menschen mit typischerweise vielen Erkrankungsdiagnosen das Risiko kaum verändert. Hiervon scheinen die Regressions- und ML-Modelle gleichermaßen betroffen zu sein. Mit Blick auf die weiteren Subgruppenmerkmale konnte festgestellt werden, dass die Prognosegüte bei Notfallaufnahmen ins Krankenhaus geringfügig schlechter ausfiel als für Normalfälle. Das Geschlecht hatte – zumindest bei der Vorhersage der Mortalität – einen geringfügigen Einfluss zugunsten der Frauen. Es gab hierbei aber keinen wesentlichen Unterschied zwischen logistischer Regression und AdaBoost.

Unterschiede zwischen den Verfahren waren im direkten Vergleich der jeweiligen Score-Ränge ersichtlich, vor allem für das Outcome Mortalität. Hier lieferte AdaBoost für eine Gruppe an Versicherten adäquatere Score-Werte als die logistische Regression. Während die Versicherten bei der logistischen Regression teilweise sehr hohe Eintrittswahrscheinlichkeiten für das Outcome Mortalität aufwiesen, ordnete AdaBoost sie eher in eine niedrige Risikogruppe ein. Die Vorhersage von AdaBoost war insofern zutreffend, als dass keiner der Versicherten tatsächlich verstarb. Bei genauerer Analyse der Versicherten fiel auf, dass diese scheinbar zu einem großen Teil einer spezifische Subgruppe von Frauen im Alter um die 30-40 Jahre mit schwangerschaftsassoziierten Diagnosen angehörten. Da die schwangerschaftsassoziierten Diagnosen im logistischen Regressionsmodell nicht enthalten sind, ist anzunehmen, dass diese Diagnosen bei AdaBoost protektiv wirken, d. h., dass Personen mit diesen Diagnosen von AdaBoost ein geringeres Risiko für das Eintreten von Mortalität zugewiesen wird. Gerade bei Schwangeren lässt sich vermuten, dass hier andere Merkmalsausprägungen häufiger dokumentiert sind, die eigentlich das Risiko zu versterben, erhöhen (z. B. multiple oder lange vorangegangene Krankenhausaufenthalte). Diese sind aber durch die besondere Situation der Schwangerschaft bedingt und in der Regel nicht mit einem erhöhten Sterberisiko assoziiert.

9 Nutzung der Prädiktion in der Routineversorgung

Im Folgenden wird auf die möglichen Anwendungsfälle für KI-Verfahren in der Routineversorgung, insbesondere in Bezug auf Routinedaten der Krankenkassen, eingegangen. Zudem werden die Voraussetzungen für die Implementierung von KI-Verfahren in Krankenkassen erläutert. Insbesondere rechtliche Grundlagen für die Datennutzung und -verarbeitung sind zu berücksichtigen, aber auch betriebswirtschaftliche Überlegungen spielen eine Rolle bei der Frage, ob entsprechende Verfahren eingesetzt werden sollen.

9.1 Einleitung

Mit der Beschreibung der verschiedenen in diesem Projekt verwendeten Routinedaten wurde bereits im ersten Kapitel dieses Weißbuchs deutlich, wie umfangreich und vielfältig das Datenvolumen ist, über das die gesetzlichen Krankenkassen verfügen. Dies gilt insbesondere für das GKV-Gesamtsystem, das rund 75 Mio. Menschen in Deutschland (ca. 90 % der Bevölkerung) repräsentiert (BMG 2024). Aber auch jede einzelne Krankenkasse erhält bzw. erzeugt erhebliche Mengen an Daten, die innerhalb der jeweiligen Organisation für verschiedene Zwecke genutzt werden können oder, unter bestimmten Voraussetzungen, das Potenzial dazu haben.

Ein Antrieb für das Nutzbarmachen von Daten könnte beispielsweise darin liegen, in der Versorgung höhere Qualität durch bessere Entscheidungen zu erreichen. Im Allgemeinen kann auch im Gesundheitswesen grob nach Ergebnis-, Prozess- und Strukturqualität (Donabedian 2005) unterschieden werden. Gute Krankenkassen sehen sich als Partner ihrer Versicherten und in „guter Qualität“ einen Vorteil im Wettbewerb mit anderen Anbietern (Knieps et al., 2023). Aber auch der Gesetzgeber verpflichtet die GKV zu entsprechenden Anstrengungen, indem er ihnen vorschreibt, dass Qualität und Wirksamkeit der Leistungen „dem allgemeinen Stand der medizinischen Erkenntnisse zu entsprechen und den medizinischen Fortschritt zu berücksichtigen“ haben (§ 2 Abs. 1 S. 3 SGB V). Durch gezielte Analysen lassen sich zudem Möglichkeiten einer effizienteren Gesundheitsversorgung identifizieren, so dass frei werdende Mittel zur weiteren Steigerung der Versorgungsqualität an anderer Stelle eingesetzt werden können.

Krankenkassen dürfen Routinedaten für eine Vielfalt an Aufgaben nutzen und unterstützen auch, etwa im Rahmen von Innovationsfondsprojekten, ihre Verwendung in der medizinischen Versorgungforschung. Neue Verfahren des maschinellen Lernens, wie sie in diesem Weißbuch vorgestellt werden, könnten in manchen Bereichen dabei helfen, präzisiere Berechnungen durchzuführen und so die Aufgabe der gesetzlichen Krankenversicherung – Erhaltung, Wiederherstellung und Verbesserung der Gesundheit der Versicherten (vgl. § 1 SGB V) – effektiver und effizienter zu erfüllen, als dies bislang möglich war. Der Schritt in Richtung dieser neuen Verfahren stellt die Krankenkassen jedoch vor verschiedene Herausforderungen hinsichtlich Implementierung und Anwendung.

Auf den nachfolgenden Seiten werden diese Herausforderungen diskutiert und mögliche Anwendungsfelder für den Einsatz von KI-Verfahren mit GKV-Routinedaten in Krankenkassen bzw. für die Zusammenarbeit mit Krankenkassen vorgestellt.

9.2 Anwendungsfälle für KI-Verfahren

Ermittlung von Krankheits- bzw. Patientenverläufen durch überwachtes Lernen

Der Schwerpunkt in diesem Weißbuch liegt auf Methoden des überwachten Lernens. Aus den für das Training der Modelle verwendeten historischen Falldaten ist direkt ablesbar, welche Bedarfe jeweils im Nachgang entstanden sind. So wissen wir z. B., dass eine zufällig ausgewählte weibliche Person im Alter von 81 Jahren nach einer Hüftoperation eine Rehamaßnahme von ihrer Krankenkasse erhalten hat oder dass sie nach 30 Tagen ungeplant erneut ins Krankenhaus aufgenommen werden musste. Es ist also bekannt, welcher „Input" zu welchem „Output" geführt hat. Die Daten sind „annotiert" (siehe Kapitel 2.1.1 zum Überwachten Lernen).

Mit der Vorhersage poststationärer Ereignisse wird im Projekt KI-THRUST das Fallbeispiel des Entlassmanagements adressiert. Neben Geschlecht und Alter sind für die Ermittlung von Bedarfen im Entlassmanagement weitere Variablen relevant. Empfehlungen hierzu lassen sich dem Expertenstandard Entlassungsmanagement in der Pflege (DNQP, 2009) entnehmen. Darunter zählen etwa frühere Krankenhausaufenthalte, eine bestehende Pflegebedürftigkeit oder eine bestehende Hilfsmittelversorgung. Solche Kriterien werden in den in KI-THRUST untersuchten Verfahren abgebildet. Das Ziel ist also, die Leistungsfähigkeit von Prognoseverfahren in einem bestehenden, fachlich gut beschriebenen Prozess mithilfe von KI zu steigern. Methoden des überwachten Lernens könnten auch für andere Handlungsfelder überprüft werden, in denen ähnliche Leitfäden oder etablierte Prozesse bestehen oder für die zumindest eine belastbare Theorie über den Zusammenhang zwischen Input- und Outputvariablen aufgestellt wurde.

Denkbar sind auch Methoden des unüberwachten Lernens. Aufgrund der Menge an unterschiedlichen Daten, die den Krankenkassen vorliegen, könnte auch die rein explorative Suche nach Mustern und Beziehungen wertvolle Erkenntnisse zutage fördern. Mit Blick auf verschiedene Regionen oder Lebenswelten könnte dies auch für kleinere oder nicht geöffnete Krankenkassen interessant sein, um Versorgungskonzepte zu entwickeln, die genau auf den eigenen Versichertenkreis angepasst sind.

Vorhersage und Vermeidung von unerwünschten Ereignissen/Krankheitsverläufen

Ein naheliegender Anwendungsfall für prädiktive KI ist Prävention, also die Vermeidung von Krankheit bzw. der Verschlechterung einer bestehenden Krankheit. Präventionsangebote sind eine wichtige Form der Versorgungsgestaltung, da erstens jede vermiedene Krankheit die Vermeidung von individuellem körperlichem oder seelischem Leid bedeutet und zweitens die individuellen und gesellschaftlichen bzw. solidargemeinschaftlichen Kosten für Behandlungen oder auch für Krankengeld bei längerer Arbeitsunfähigkeit gar nicht erst entstehen.

Aus Routinedaten können mithilfe von Prädiktionsverfahren Krankheitsverläufe und damit einhergehende Versorgungsbedarfe vorhergesagt werden. Daraus lassen sich Präventionsmaßnahmen mit dem Ziel ableiten, Wahrscheinlichkeiten für unerwünschte Verläufe zu verringern. Präventionsmaßnahmen zielen z. B. auf Bewegungsgewohnheiten, Stressmanagement, Ernährung oder etwaigen Suchtmittelkonsum ab. Mit KI-Verfahren könnten die hierfür infrage kommenden Versicherten möglicherweise treffsicherer ermittelt und Präventionsangebote zugleich weiter ausdifferenziert werden. Unabhängig von verfügbaren Präventionsangeboten könnten Risiken für schwere Erkrankungen aller Art vorhergesagt und den Versicherten damit der Gang zum Haus- oder Facharzt nahegelegt werden. Der Gesetzgeber hat hierzu neue Rechtsnormen geschaffen, die weiter unten dargestellt werden.

Auch bei größten präventiven Anstrengungen lassen sich viele Behandlungen nicht vermeiden. In solchen Fällen kann jedoch eine präzise Vorhersage den behandelnden Leistungserbringern und Verwaltungsmitarbeitenden dabei helfen, den Patientenverlauf möglichst reibungslos zu gestalten.

Im Krankenhausentlassmanagement nach § 39 Abs. 1a SGB V wird diese Herausforderung besonders deutlich. Damit keine Versorgungsbrüche entstehen, sollen die behandelnden Krankenhäuser so früh wie möglich die Nachsorge ihrer Patientinnen und Patienten organisieren. In vielen Fällen ist der Bedarf jedoch zunächst nicht eindeutig und aufgrund fehlender Gesundheitsdaten oder Verständigungsschwierigkeiten werden Bedarfe manchmal erst spät oder gar nicht erkannt. Folgen können eine verzögerte Fortsetzung der Behandlung und ungeplante Wiederaufnahmen sein (Rageth et al., 2023).

Eine verbesserte Prognose von Entlassmanagementbedarfen war das zentrale Ziel des Innovationsfondsprojekts USER (Umsetzung eines strukturierten Entlassmanagements mit Routinedaten). Hier implementierten die Projektbeteiligten einen Prognosealgorithmus – in Form einer logistischen Regressionsgleichung – über eine neue Schnittstelle zwischen mehreren Krankenkassen und Krankenhäusern. Auf diesem Wege konnten zeitnah Prognosescores ermittelt und den Krankenhausmitarbeitenden in einer Software bereitgestellt werden, damit ggf. erforderliche Maßnahmen im Rahmen des Entlassmanagements frühzeitig veranlasst werden konnten. Mit dieser Maßnahme konnte die Zahl der ungeplanten Wiederaufnahmen im Projektzeitraum um 13 % gesenkt werden (Broge et al., 2024). Ausgangspunkt des Nachfolgeprojekts KI-THRUST war die Frage, ob sich die Prognosegenauigkeit im Rahmen eines entsprechenden Versorgungskonzepts oder bei vergleichbaren Vorhaben durch den Einsatz von KI-gestützten Methoden steigern lässt.

9.3 Voraussetzungen für die Implementierung von KI-Verfahren mit Routinedaten der Krankenkassen

9.3.1 Datenverfügbarkeit

Für das Training von KI-Modellen sind Daten erforderlich. Während der Instrumentenkasten der klassischen Statistik auch Lösungen bereithält, die verwendet werden können, um für Datensätze mit wenigen Beobachtungen wissenschaftliche Aussagen zu treffen, sind maschinelle Lernverfahren erst bei großen Datenvolumen sinnvoll einsetzbar bzw. ist es umgekehrt erst mit solchen Verfahren möglich, extrem große Datenmengen (Stichwort „Big Data“) zu bewältigen.

Krankenkassen verfügen über sehr umfangreiche gesundheitsbezogene Daten, die vorrangig zum Zweck der Abrechnung von Versorgungsleistungen regelmäßig anfallen. Sie werden als Routinedaten bezeichnet und sind so umfangreich und vielfältig, dass ihnen ein hohes Zweitverwertungspotenzial zugeschrieben werden kann. Die im Projekt betrachteten Routinedaten sind ausführlich in Kapitel 1 beschrieben, womit Einschätzungen auch zu allgemeinen Nutzungsmöglichkeiten von Routinedaten bei Krankenkassen über die hier näher beschriebenen Anwendungen hinaus möglich sein sollten.

9.3.2 Rechtsgrundlagen für Innovationen und anwendbare Verfahren für die Datennutzung

Bei den hier betrachteten Routinedaten der Krankenkassen handelt es sich um Sozialdaten, die z. B. Aufschluss über die Gesundheit und weitere persönliche Merkmale einzelner Personen geben können. Sie unterliegen grundsätzlich dem Sozialgeheimnis (§ 35 SGB I) und dürfen nicht „unbefugt verarbeitet werden“.

Aufgrund der hohen Anforderungen an den Schutz der Daten, müssen Forschende sich in diesem Bereich fortbilden und ihren Wissensstand aktuell halten. Dazu zählt auch die europäische Gesetzgebung. Beispielhaft genannt seien hier die Datenschutzgrundverordnung (DSGVO) und die Verordnung über künstliche Intelligenz (KI-VO), die bei der Nutzung von GKV-Routinedaten und maschinellen Lernverfahren besonders im Fokus stehen.

Wer Angebote oder Forschungsvorhaben aufgrund einer Datenauswertung von Routinedaten plant, muss sich umfassend mit den hierfür in Frage kommenden Rechtsgrundlagen vertraut machen. Viele Vorhaben müssen bei der jeweils zuständigen Landes- oder Bundesaufsicht angezeigt bzw. von dieser genehmigt werden. Im Rahmen dieses Weißbuchs ist hierzu keine hinreichende Betrachtung möglich. Die folgenden Beispiele können aber einen Einstieg in die Thematik bieten, insbesondere für Forschende, die mit einem Forschungsvorhaben an Krankenkassen herantreten möchten. Je nach Aufgabenstellung sind weitere, hier nicht genannte Rechtsgrundlagen erforderlich.

Für den Einstieg in die Recherche sei zunächst auf die §§ 284-287 SGB V verwiesen, in denen die „Grundsätze der Datenverarbeitung“ bestimmt werden. § 284 SGB V zählt abschließend auf, für welche Aufgaben die Krankenkassen personenbezogene Daten ihrer Versicherten erheben dürfen.

Hieran schließen sich weitere Rechtgrundlagen an, die für spezifische Anwendungsbereiche eingeführt wurden. Die Bundesregierung nennt unter Verweis auf den Bericht des GKV-Spitzenverbands zum Stand der Förderung von Versorgungsinnovationen im Jahr 2023 gemäß § 68b Absatz 4 SGB V an das Bundesministerium für Gesundheit beispielhaft Rechtsgrundlagen für Versorgungsinnovationen (Deutscher Bundestag 2024):

- § 140a SGB V (Besondere Versorgung)

- § 64 ff. SGB V (Modellvorhaben)
- § 20 ff. SGB V (Prävention und Gesundheitsförderung)
- § 92a SGB V (Innovationsfonds)
- § 73b SGB V (Hausarztzentrierte Versorgung)
- § 43 SGB V (Ergänzende Leistungen zur Rehabilitation)

Als Rechtsgrundlage für die Datenübermittlung im Innovationsfondsprojekt KI-THRUST kam § 92a SGB V in Verbindung mit § 75 SGB X zum Tragen.

Übermittlung von Sozialdaten für die Forschung nach § 75 SGB X

Zum Zweck der Versorgungsforschung können Krankenkassen und andere Sozialversicherungsträger Sozialdaten von Versicherten an Dritte übermitteln, etwa an wissenschaftliche Forschungseinrichtungen. Erforderlich hierfür ist eine Genehmigung nach § 75 SGB X durch die zuständige Aufsichtsbehörde.

Für das Genehmigungsverfahren für die Datenübermittlung durch bundesunmittelbare gesetzliche Krankenkassen stellt das Bundesamt für Soziale Sicherung (BAS) ein regelmäßig aktualisiertes Antragsformular[12] zur Verfügung, in dem Fragen z. B. zum Inhalt des Vorhabens, zu den Sendern und Empfängern der Daten oder zu Art und Umfang der Daten beantwortet werden müssen.

Der Aufsichtsbehörde ist stets auch ein Datenschutzkonzept vorzulegen (§ 75 Abs. 1 Satz 4 SGB X). Das in KI-THRUST vom aQua-Institut erstellte Datenschutzkonzept beinhaltet u. a. eine Datensatzbeschreibung, die grafische Darstellung der Datenflüsse, eine Beschreibung der datenverarbeitenden Prozesse und Verantwortlichkeiten, technisch-organisatorischen Maßnahmen (TOM) zur Einhaltung des Datenschutzes sowie Löschfristen.

In KI-THRUST wurde beispielsweise der Antrag von der Konsortialführung selbst, dem aQua-Institut, vorbereitet und eingereicht, die damit auch als zentraler Ansprechpartner für das BAS fungierte. Die kooperierenden Betriebskrankenkassen schlossen sich nach Prüfung aller Unterlagen diesem Antrag an, indem sie eine standardisierte Erklärung („One Pager“) abgaben, die dem Antrag beigelegt wurde. Nachdem das BAS den Antrag genehmigt hatte, konnten die Kooperationspartner die Sozialdaten gemäß Datensatzbeschreibung an das aQua-Institut übermitteln.

Neben dem Formular für den Hauptantrag hat das BAS auch für die „One Pager“ sowie diverse weitere einzureichende Dokumente Musteranlagen entwickelt, die es in der jeweils aktuellen Version bereitstellt.

Da jedes Projekt inhaltlich einzigartig ist, die gesetzlichen Regeln einem ständigen Veränderungsprozess unterliegen und die Beteiligung landesunmittelbarer Kostenträger länderspezifische Besonderheiten erforderlich machen kann – der Antrag muss in solchen Fällen (zusätzlich) bei der jeweils zuständigen Landesaufsicht gestellt werden – sollte dieser Prozess so weit wie möglich im Voraus geplant werden.

Auswertungen zur individuellen Ansprache von Versicherten

Seit dem Inkrafttreten des Gesetzes für eine bessere Versorgung durch Digitalisierung und Innovation (Digitale-Versorgung-Gesetz – DVG) vom Dezember 2019 können Krankenkassen auch mittels § 68b SGB V Versorgungsinnovationen fördern. Diese Innovationen sollen „insbesondere ermöglichen, 1. die Versorgung der Versicherten anhand des Bedarfs, der aufgrund der Datenauswertung ermittelt worden ist, weiterzuentwickeln und 2. Verträge mit Leistungserbringern unter Berücksichtigung der Erkenntnisse nach Nummer 1 abzuschließen“. Die Krankenkasse darf die Daten pseudonymisiert bzw. anonymisiert auswerten, um darauf basierend neue Angebote bzw. Verträge zu entwickeln, um dann

12 abrufbar unter: https://www.bundesamtsozialesicherung.de/de/service/rundschreiben/detail/antraege-auf-verarbeitung-von-sozialdaten-fuer-die-forschung-und-planung-im-sozialleistungsbereich-gemaess-75-sgb-x/, Stand 02/2025

ihren Versicherten „individuell geeignete Versorgungsinnovationen oder sonstige individuell geeignete Versorgungsleistungen“ anzubieten. Mit den Angeboten darf die Kasse nicht in die Therapiefreiheit oder die Wahlfreiheit der Versicherten eingreifen und die Versicherten können jederzeit der Unterbreitung solcher Angebote widersprechen.

Mit dem 2024 durch das Gesundheitsdatennutzungsgesetz (GDNG) eingeführten § 25b SGB V eröffnen sich weitere Möglichkeiten für die Krankenkassen, die Daten ihrer Versicherten zu deren Gunsten zu nutzen. Demnach können Krankenkassen nun „zum Gesundheitsschutz eines Versicherten datengestützte Auswertungen vornehmen und den Versicherten auf die Ergebnisse dieser Auswertung hinweisen“. Dies ist derzeit beschränkt auf die Erkennung von

- seltenen Erkrankungen,
- Krebserkrankungen,
- schwerwiegenden Gesundheitsgefährdungen durch eine Arzneimitteltherapie,
- Pflegebedürftigkeit,
- ähnlich schwerwiegenden Erkrankungen (soweit dies aus Sicht der Kassen im überwiegenden Interesse der Versicherten ist) und
- Impfindikationen (soweit von der Ständigen Impfkommission empfohlen).

Stellt die Kasse bei ihrer Auswertung eine der o. g. Gefährdungen, Erkrankungsrisiken oder Impfindikationen fest, erhält der Versicherte von ihr eine begründete Empfehlung, sich ärztlich, pflegerisch o. ä. beraten zu lassen. Analog zur Regelung des § 68b SGB V bleiben ärztliche Therapiefreiheit und Wahlfreiheit der Versicherten unangetastet. Versicherte können die Empfehlungen ignorieren und den Auswertungen widersprechen, ohne dass ihnen dadurch Nachteile entstehen dürfen.

9.3.3 Ressourcen

Die Aufbereitung und die Verarbeitung von Daten erfordern personelle und finanzielle Aufwände. Zum relevanten Fachpersonal zählen insbesondere Datenanalysten und Data Scientists, Informatiker sowie Rechtswissenschaftler mit Schwerpunkt Datenschutz. Diese müssen in der Lage sein, zunehmend komplexe Verfahren vorzubereiten und durchführen zu können.

Neben der Beschäftigung von Fachpersonal sind angemessene technische Voraussetzungen zu schaffen. Da KI-Verfahren ihr ganzes Potenzial erst ab einer bestimmten Menge an Daten und Anzahl an Variablen entfalten, müssen Krankenkassen auch hier mit Investitionen und Upgrades sowie regelmäßigem Nach- bzw. Neutrainieren der Modelle rechnen.

Der Einstieg hinsichtlich Software ist zwar oft günstig, da die gängigen Programmiersprachen (bei KI-THRUST: R und Python) und Pakete meist quelloffen und kostenlos verfügbar sind. Mitarbeitende müssen jedoch regelmäßig geschult und die Installationen müssen auf dem aktuellen Stand gehalten werden. Sobald Themen in größeren Teams bearbeitet und neue Prozesse fest in der Organisation implementiert werden sollen, können Kosten für Anwendungspakete und Beratung entstehen.

Je nach Anwendungsfall können Krankenkassen, ähnlich wie es bei Standardsoftware der Fall ist, auch im KI-Bereich auf vorgefertigte Lösungen zurückgreifen, die sie entweder selbst implementieren oder extern verwalten lassen.

Extern Beauftragte könnten auch zentrale Funktionen bei der Umsetzung des Konzepts des so genannten *federated learning* (föderiertes Lernen) übernehmen. Krankenkassen könnten hierbei lokale Modelle mit ihren eigenen Datensätzen trainieren. Der beauftragte Dritte erhält dann nicht die eigentlichen (Sozial-) Daten, sondern nur die Parameter der Einzelmodelle, die er in einem globalen Modell integriert (McMahan et al., 2016). Dieses könnte dann wiederum von den einzelnen Kassen auf die jeweiligen Anforderungen hin angepasst und verwendet werden.

Ist ein neues Verfahren einmal eingerichtet, muss es kontinuierlich an die aktuellen technischen und rechtlichen Erfordernisse angepasst werden. Ein Nachtrainieren bzw. Neutrainieren der Modelle auf aktuellen Daten ist notwendig, um stets repräsentative und damit zielführende Ergebnisse zu erhalten. Zudem kann es sinnvoll sein, Erkenntnisse aus der Nutzung der Modelle zurückfließen zu lassen, beispielsweise in Form von systematisch erhobenem Feedback von Versicherten zu Verfahren, in denen KI eingesetzt wurde. Je nach Anwendungsfall ist zu prüfen, ob eigenverantwortliche Anpassungen unter den jeweiligen rechtlichen Bedingungen zulässig sind oder erst von einer anderen Stelle genehmigt werden müssen.

9.3.4 Strategie

Die Implementierung von KI-Verfahren beschränkt sich nicht auf technische und juristische Aspekte, sondern ist auch für Krankenkassen eine ganzheitliche Managementaufgabe. Recht, Technik und Management gehen bei Einführung und Betrieb von KI-Verfahren Hand in Hand. Aus den Unternehmenszielen leitet sich der Bedarf an KI-Anwendungsfällen ab, woraus sich der Bedarf an Technik und Personal ergibt, welches die Verfahren unter Berücksichtigung der rechtlichen Vorgaben umsetzt.

Geeignete Anwendungsfälle für KI-Verfahren finden sich dort, wo klare und messbare Ziele vorgegeben werden können, z. B. die quantitative Verbesserung von medizinisch beobachtbaren Outcomes. So wurde bei KI-THRUST der Einfluss verschiedener Variablen auf die Outcomes „Mortalität“ und „Ungeplante Wiederaufnahmen“ nach einer vorangegangenen Entlassung aus einem Krankenhaus geprüft. Ein konkretes Ziel könnte also sein, durch den Einsatz verbesserter Prognosen den Anteil ungeplanter Wiederaufnahmen um einen konkreten Prozentsatz innerhalb eines Jahres zu verringern. Aber auch qualitative Maßnahmen wie die Verbesserung von Verwaltungsabläufen mit KI oder die Einführung KI-gestützter neuer Beratungsleistungen können mithilfe von Mitarbeitenden- bzw. Kundenbefragungen in messbare Ziele übersetzt werden.

Auch bei der Einführung von KI-Verfahren in bestehende Strukturen gilt: Je schneller und klarer sich ein Erfolg einstellt ist, desto leichter lassen sich die Beteiligten von der Veränderung überzeugen. Deshalb erscheint es als empfehlenswert, den Fokus zunächst auf die sichtbare Verbesserung einzelner Prozesse zu setzen, bevor Organisationsstrukturen grundlegend verändert werden (vgl. Dorninger et al., 2024). Alle betroffenen Mitarbeitenden werden bestenfalls bereits bei der Entwicklung des Prozesses eingebunden. Dies senkt das Risiko, Lösungen zu entwickeln, die im Alltag nicht oder fehlerhaft angewendet werden, weil sie nicht zu den eingeübten Arbeitsabläufen passen.

Zu den Beteiligten zählen neben Krankenkassen und ihren Versicherten und Mitarbeitenden insbesondere die Leistungserbringer. Sollte beispielsweise auf Basis von KI-THRUST ein Unterstützungssystem wie im Projekt USER im Versorgungsalltag implementiert werden, sind Verträge zwischen Kassen, Krankenhäusern und Softwareherstellern zu schließen. Sie sind daher frühestmöglich einzubinden. Eine Rückmeldung des Krankenhauspersonals aus USER war beispielsweise, dass die Darstellung der Prognoseergebnisse nicht direkt in das Krankenhausinformationssystem implementiert werden konnte. Der parallele Betrieb der zusätzlichen Software war zwar technisch problemlos, bedeutete aber einen Bruch im Bedienkomfort (vgl. Broge et al., 2024). Solche Herausforderungen sind gemeinsam zu erörtern.

Eine weitere Herausforderung, die alle Beteiligten von KI-Prozessen betrifft, ist die Komplexität der Berechnungswege von KI-Verfahren. Wie ein Ergebnis zustande gekommen ist, lässt sich oft nur unter Zuhilfenahme weiterer statistischer Methoden erklären (vgl. Kapitel 7.2). Dies erfordert sowohl hohe eigene Kompetenz als auch ein gewisses Maß an Vertrauen in die Richtigkeit der Ergebnisse. Fühlen sich Versicherte bzw. Patienten übergangen, leidet das Vertrauensverhältnis und sinnvolle Angebote werden möglicherweise abgelehnt. Umgekehrt könnten progressive Versicherte eine Kasse verlassen, wenn diese weitgehend auf KI-Lösungen verzichtet und darum als zu wenig innovativ wahrgenommen wird.

Kundenberaterinnen und -berater in Krankenkassen, oder auch Leistungserbringende, die von der Notwendigkeit des Einsatzes von KI überzeugt sind und die Ergebnisse für die von Ihnen betreuten Versicherten bzw. Patienten laienverständlich plausibilisieren können, werden KI-Verfahren eher einsetzen als solche, die hierzu noch keinen Bezug aufbauen konnten.

Ist die Einführung eines neuen (KI-)Verfahrens zunächst erfolgreich, ist dies noch keine Garantie für eine dauerhafte Verbesserung. Ähnlich wie in o. g. technischer Hinsicht müssen die Verfahren auch auf strategischer Ebene gepflegt werden, indem eine systematische Evaluation und Weiterentwicklung bzw. Anpassung im Rahmen der Unternehmensentwicklung stattfindet. Damit unterscheiden sie sich nicht von anderen Prozessen innerhalb einer Unternehmung, die regelmäßig Aktualisierungsmethoden wie dem PDCA-Zyklus („Plan, do, check, act“) unterzogen werden müssen.

9.4 Fazit

KI-Verfahren können in Krankenkassen an verschiedenen Stellen eine große Rolle spielen. In allgemeinen Geschäftsprozessen, z. B. der Kundenberatung oder im Marketing, wird KI künftig ebenso eine Rolle spielen, wie in Unternehmen anderer Branchen auch. Chatbots helfen bereits heute, Versichertenfragen zu beantworten und Abrechnungen werden mit KI auf ihre potentielle Fehlerhaftigkeit untersucht.

Die Besonderheit bei Krankenkassen besteht in diesem Zusammenhang darin, dass sie Sozialdaten erheben und speichern. Diese unterliegen einem besonderen Schutz, mit dem sich auch Forschende auseinandersetzen müssen, wenn sie eine Zusammenarbeit mit Krankenkassen planen. Dem damit verbundenen Aufwand steht die Chance gegenüber, aus diesen Daten mit den richtigen Algorithmen möglicherweise Erkenntnisse abzuleiten, die zu besseren Entscheidungen für Gesundheit und Wohlbefinden führen könnten. KI-Verfahren könnten hier in manchen Fällen zu besseren Prädiktionen kommen als herkömmliche Verfahren und so für eine Verbesserung der Gesundheitsentscheidungen sowohl für einzelne Versicherte als auch für die gesamte Bevölkerung sorgen.

Ein zentraler Aspekt betrifft die Erklärbarkeit bzw. Interpretierbarkeit von KI-Modellen. Die Endverantwortung für eine Entscheidung hat immer ein Mensch zu tragen. Gerade für gesundheitsbezogene Anwendungsfälle erscheint dies relevant. Für den Erfolg der Einführung von KI-Verfahren ist daher ausschlaggebend, dass alle Beteiligten – z. B. Mitarbeitende, Führungskräfte, Versicherte, Leistungserbringer – hinter den Prozessen stehen können.

Ausschlaggebend ist auch, dass ein neues Verfahren auch eine tatsächliche Verbesserung bringt. In Kapitel 6 wurde gezeigt, dass in unserem spezifischen Anwendungsfall „Entlassmanagementbedarf“ nicht mit jedem KI-Verfahren präzisere Prognosen als mit der logistischen Regression erzielt wurden. Je nach Aufgabenstellung können daher unterschiedliche Ansätze vorzuziehen sein. Für jeden Anwendungsfall ist auch zu überprüfen, ob die erzielte Verbesserung des Outcomes tatsächlich den Aufwand und die Kosten der Prozessanpassung übersteigt. Diese beinhalten nicht nur Investitionen in Personal und Ausstattung, sondern auch Aufwendungen für Öffentlichkeitsarbeit, rechtliche Prüfungen, Schulungsbedarfe u.v.m. Auf der Outcome-Seite müssen auch die Präferenzen der Patientinnen und Patienten bzw. Versicherten einbezogen und das neue Verfahren regelmäßig evaluiert werden. Gerade zu Beginn sollte der Zusatznutzen klar erkennbar sein, um auch kritische Stakeholder vom Sinn einer KI-Nutzung zu überzeugen.

Wie andere Organisationen sind auch Krankenkassen gezwungen, sich mit dem Thema KI auseinanderzusetzen und aktiv zu entscheiden, für welche Prozesse sie die neuen Verfahren einsetzen sollen und an welcher Stelle sie ihnen keinen positiven Nutzen zuordnen. Dass neue und umfangreiche Rechtsgrundlagen, darunter nicht zuletzt die KI-Verordnung der EU, dabei zu berücksichtigen sind, erschwert den Einstieg in die Materie. Dieser gelingt, wenn allein oder in Partnerschaft ausreichend Daten generiert, menschliches Know-How und technische Ressourcen verfügbar gemacht werden und die Umsetzungsstrategie erfolgreich ist.

Quellen

Broge, B., Lingnau, R., Pollmann, T., Willms, Dr.G. (alle aQua-Institut), Blum, Dr. K. (DKI) (2024). Ergebnisbericht des Projekts „Umsetzung eines strukturierten Entlassmanagements mit Routinedaten“.

Bundesministerium für Gesundheit (BMG) (2024): *Ergebnisse der GKV-Statistik KM1, Stand: 30. Dezember 2024*.

Deutscher Bundestag (2024): Drucksache 20/12688 (21.8.2024)

Deutsches Netzwerk für Qualitätsentwicklung in der Pflege (Hrsg.) (DNQP) (2009). Expertenstandard Entlassungsmanagement in der Pflege.

Donabedian, A. (2005). Evaluating the Quality of Medical Care. In: The Milbank Quarterly, 83(4), 691-729.

Dorninger H., Klar A., Storms K., Tremp K., Wilhelm D., Gliwa J., Benn A., Schlegel C. (2024): A (Gen)AI Pathfinder. Unlock the Potenzial of (Gen)AI in health insurance. Boston Consulting Group.

Knieps F., Klemm A.-K., Demmler G (Hrsg.) (2023): BKK Kundenreport 2023 - Qualität von Krankenkassen. Fokus Nachhaltigkeit.

McMahan H. B., Moore E., Ramage D., Hampson S., Aguera y Arcas, B. (2017): Communication-Efficient Learning of Deep Networks from Decentralized Data. International Conference on Artificial Intelligence and Statistics, 1273–1282. http://proceedings.mlr.press/v54/mcmahan17a/mcmahan17a.pdf

Rageth, L., Leuppi, J. D., Leuppi-Taegtmeyer, A. B., Lüthi-Corridori, G., Boesing, M. (2023). Prädiktoren für frühe Ungeplante Rehospitalisationen. Praxis 2023; 112 (2): 75–81.

10 Zusammenfassung

10.1 Ergebnisse der Routinedatenanalysen

- **Vergleich der Ergebnisse zu unterschiedlichen Outcomes:** Die im Projekt exemplarisch betrachteten zwei Outcomes Mortalität[13] und Ungeplante Wiederaufnahme[14] lassen sich mit den eingesetzten Machine Learning (ML)-Verfahren[15] unterschiedlich gut vorhersagen. Anhand etablierter Evaluationsmetriken (AUC-ROC) lässt sich für das Outcome Mortalität feststellen, dass alle Modelle formal eine ausgezeichnete Vorhersagegüte (AUC-ROC > 0,8) erreichen. Im Gegensatz dazu liegt die Vorhersagequalität für das Outcome Ungeplante Wiederaufnahme bei allen Verfahren formal unter einer gemeinhin als akzeptabel angenommenen Schwelle (AUC-ROC < 0,7) Das Outcome Mortalität ist demnach deutlich besser anhand der hier genutzten Routinedaten der Krankenkassen vorhersagbar als das Outcome Ungeplante Wiederaufnahme. Bei entsprechenden Aussagen ist jedoch stets zu beachten, dass AUC-ROC-Werte als losgelöst betrachtete Kennzahl kaum Rückschlüsse auf die Sinnhaftigkeit des Einsatzes eine Vorhersagemodells in der Praxis erlauben. Diese wird maßgeblich auch von anderen Faktoren mitbestimmt, z.B. von der erwartbaren Häufigkeit von Outcomes in der Population, für die Vorhersagen ermittelt werden sollen und von den Konsequenzen bzw. Interventionen, die dann bei bestimmten Vorhersagen geplant sind. Deutlichere Hinweise zur Praktikabilität von Vorhersagen als AUC-ROC-Werte können AUC-PR-Werte bzw. Precision-Recall-Kurven liefern, was durch ihre Abhängigkeit von der Prävalenz der Outcomes mitbedingt ist (weshalb sie dann allerdings zugleich für Vergleiche von Modellierungen zu unterschiedlichen Outcomes weniger geeignet sind). Entsprechende Ergebnisse zeigen, dass die Präzision der Modelle für beide Outcomes eher gering ausfällt, also auch Personen aus „Risikogruppen“ mit vergleichsweise hohen vorhergesagten Risiken nachfolgend häufig real nicht von den jeweils betrachteten Outcomes betroffen sind. Ein wesentlicher Grund hierfür liegt in der starken Unbalanciertheit des Datensatzes bzw. der niedrigen Prävalenz der positiven Klassen, die auch in vielen epidemiologischen Studien dazu führt, dass sich zwar Risikofaktoren bzw. Merkmalsausprägungen identifizieren lassen, die mit teils stark erhöhten relativen Risiken assoziiert sind, aber auch Personen mit diesen Risikofaktoren innerhalb begrenzter Zeiträume eher selten betroffen sind (beispielhaft und plakativ: Auch Kettenraucher überleben i.d.R. mehrheitlich die kommenden 365 Tage). Was mit diesen Überlegungen

[13] Definition d. Outcomes „Mortalität“: Versterben innerhalb von 30 Tagen nach Entlassung
[14] Definition d. Outcomes „Ungeplante Wiederaufnahme“: Stationäre Wiederaufnahme mit Aufnahmegrund „Notfall“ innerhalb von 30 Tagen nach Entlassung
[15] Verwendete Verfahren: Logistische Regression, Random Forest, Adaptive Boosting (kurz AdaBoost), Neuronales Netz

deutlich werden sollte ist, dass auch die Präzision eines Vorhersagemodells ein Kennwert ist, der nur im Kontext mit weiteren Informationen und Abwägungen zu Entscheidungen über potenziell sinnhafte Nutzungen von Prädiktionsmodellen beitragen kann.

- **Vergleich der Ergebnisse zu Modellvarianten M1 bis M3:** Im Projekt wurden drei unterschiedlich umfangreiche Sets an Variablen bzw. Features[16] im Sinne potenzieller Prädiktoren zusammengestellt. Ziel war es, zu untersuchen, wie sich die Vorhersagegen verändern, wenn die verschiedenen Modelle auf Basis unterschiedlich umfangreicher Informationen (im Sinne von Merkmalen/Features) trainiert werden. Aus dem Vergleich der drei Modellvarianten geht hervor, dass bereits das Modell 1 (Basisprädiktoren) eine relativ hohe Erklärungskraft für das Eintreten der beiden Outcomes aufweist. Mit dem Modell 2 konnte die Vorhersagegenauigkeit erwartungsgemäß weiter verbessert werden, indem neben den Basisprädiktoren zusätzlich Vorerkrankungsdiagnosen berücksichtigt wurden. Für Modell 3 wurden die Informationen zu Vorerkrankungen zusätzlich entsprechend den Quartalen ihrer Dokumentation(en) im Vorfeld der Krankenhausbehandlungsfalls differenziert, womit sich die Vorhersagegüte der Modelle jedoch i.d.R. nicht verbesserte und z.T. auch verschlechterte. Das Modell 2 wurde vor diesem Hintergrund als das beste und somit finale Modell für die vergleichende Testung ausgewählt.

- **Vergleich der Ergebnisse zu ML-Verfahren (Kernelement des Projektes):** Die vergleichende Testung eingesetzten Vorhersageverfahren auf Basis identischer Daten zeigt, dass im zuvor beschriebenen Setting die Verfahren *AdaBoost* und *logistische Regression* die höchste Genauigkeit bei der Vorhersage der Mortalität und der Ungeplanten Wiederaufnahme erzielen. Die Unterschiede (zugunsten *AdaBoost*) sind allerdings eher gering und dürften für die praktische Umsetzung zunächst kaum Relevanz besitzen. Im Vergleich dazu schneiden *Neuronale Netze* in der hier gewählten Implementierung bei identischen Daten gemäß der drei Modellvarianten 1-3 bei beiden Outcomes schlechter ab. Hierzu ist anzumerken, dass Feature Engineering eine zentrale Rolle für die Leistungsfähigkeit von Prädiktionsmodellen spielt. Das Potenzial komplexerer Verfahren, wie *AdaBoost* oder *Neuronaler Netze,* wurde so möglicherweise noch nicht vollständig ausgeschöpft. Zukünftige Analysen sollten daher auch auf erweitertes Feature Engineering setzen. Dabei könnte ein großes Potenzial auch in der Anwendung aktueller Transformer-basierter Daten-Enkodierungen liegen, da diese in der Lage sind, komplexe Abhängigkeiten in den Daten zu erfassen.

- **Prognosegüte in Folgejahren sowie Fehlklassifikationen:** Im Projekt wurde zusätzlich untersucht, inwieweit die Vorhersagegüte der basierend auf Daten 2018 entwickelten Modelle mit *logistischer Regression* und des (besten) ML-Verfahrens *AdaBoost* bei einer Nutzung zur Vorhersage in Daten zu den beiden Folgejahren 2019 und 2021 sowie bei einer Nutzung in ausgewählten Versichertensubgruppen variiert. Die Nutzbarkeit in Folgejahren ist insofern besonders wesentlich, als dass sie dem typischen Anwendungsfall entspricht, in dem ein mit verfügbaren Daten und bekannten Outcomes entwickeltes Modell mit neu erfassten Daten genutzt wird (bei dem in der Praxis noch niemand das Outcome kennt). Gemessen an AUC-ROC-Werten erwiesen sich beide Modelle auch für das Jahr 2019 als geeignet, tendenziell lagen hier die Werte der ermittelten Gütemaße sogar noch etwas höher. Bei einer Anwendung der Vorhersagen für Behandlungsfälle im Jahr 2020 zeigten sich demgegenüber reduzierte Gütemaße, was aufgrund der gravierenden Auswirkungen der Coronapandemie auf die stationäre Versorgung den Erwartungen entspricht. Die Veränderung der Gütemaße bewegte sich jedoch auch hier nur in einem begrenzten Rahmen, womit die Nutzbarkeit der Vorhersagemodelle auch im ersten Jahr der Coronapandemie allenfalls graduell eingeschränkt war. Dies Ergebnis übertrifft anfängliche Erwartungen im positiven Sinne, da es in den Jahrzehnten vor 2020 in Deutschland kaum ähnlich gravierende Einschnitte bezüglich der gesundheitlichen Versorgung gegeben haben dürfte und insofern auch deutliche Einschränkungen bei der Prognosegüte hätten erwartet werden können.

[16] Modell 1: Basisprädiktoren (z. B. Alter, Geschlecht, Pflegegrad); Modell 2: Basisprädiktoren + Vorerkrankungsdiagnosen; Modell 3: Basisprädiktoren + quartalsabhänige Vorerkrankungsdiagnosen

- Innerhalb von Subgruppen der stationär behandelten Versicherten zeigten sich – gemessen an AUC-ROC-Werten – unterschiedliche Güten der Modellvorhersagen. Eingeschränkte Prognosemöglichkeiten zeigten sich insbesondere innerhalb der Gruppen von Personen im Alter von über 80 Jahren sowie bei Personen mit längerfristiger Unterbringung in Pflegeheimen. Die „schlechteren" Prognosemöglichkeiten innerhalb dieser Gruppen dürften maßgeblich daraus resultieren, dass innerhalb von entsprechenden „Hochrisikogruppen" naturgemäß die Varianz relevanter Risikofaktoren begrenzt ist und verbleibende Risikofaktoren die Risiken innerhalb dieser Gruppen nur noch eingeschränkt differenzieren können. Wesentlich erscheint die Beobachtung, dass Veränderungen der Vorhersagegüte in Subgruppen bei beiden Modellen (*logistische Regression* und *AdaBoost*) i.d.R. sehr ähnlich sind, womit sich keine Hinweise auf Abnormitäten der Vorhersagen in Subgruppen bei einem der beiden Modelle ergeben. Zu einer entsprechenden Schlussfolgerung führen – abgesehen von einer erklärbaren Abweichung in bestimmten Bereichen von Vorhersagewerten – auch Gegenüberstellungen von Vorhersagewerten aus beiden Modellen in Streudiagrammen, welche die Gleichartigkeit der Vorhersagen grundsätzlich sehr anschaulich belegen.

- **Erklärbarkeit:** Im Gegensatz zur vergleichsweise gut interpretierbaren *logistischen Regression* sind komplexere maschinelle Lernverfahren oft weniger transparent in Bezug auf die Bedeutung einzelner Prädiktoren (z.B. Versichertenmerkmale, Diagnosen) für die Vorhersageergebnisse. Die Relevanz einzelner Prädiktoren lässt sich bei solchen Modellen jedoch mithilfe erklärender Verfahren wie Shapley-Werte quantifizieren, die versuchen, den Beitrag eines Merkmals zur Vorherhersage zu ermitteln. Im Projekt wurden somit für die berechneten ML-Verfahren verschiedene Methoden zur Erklärbarkeit getestet, um die Bedeutung einzelner Prädiktoren bewerten zu können. Die entsprechenden Erklärbarkeitsanalysen haben gezeigt, dass Integrated Gradients im Vergleich zu Shapley Value Sampling und LIME weniger zuverlässige Relevanzzuweisungen liefert. Dies liegt vermutlich an seiner starken Abhängigkeit vom Gradienten und dem Modelloutput, die bei kleinen Werten zu kaum interpretierbaren Ergebnissen führen. Im Gegensatz dazu identifizieren LIME und Shapley konsistent relevante Merkmale, insbesondere schwerwiegende Vorerkrankungen, die plausibel mit den Outcomes Ungeplante Wiederaufnahme und Mortalität zusammenhängen. Trotz dieser ersten Einblicke bleibt die Erklärbarkeit der Modellvorhersagen begrenzt, sodass derzeit keine verlässlichen Aussagen darüber getroffen werden können, warum ein spezifischer Patient eine bestimmte Vorhersage erhält.

10.2 Nutzbarkeit von Krankenkassendaten für ML-Verfahren

- **Vorteile von Routinedaten:** Abrechnungsdaten bei Krankenkassen haben durch die gesetzlich vorgeschriebene Erfassung und Verarbeitung viele, generelle Vorteile für die (forschungsbezogene) Datenanalyse und somit auch für die Entwicklung ML-gestützter Vorhersagemodelle. So liegen die Daten in standardisierter und strukturierter Form für jeden Versicherten vor. Zudem sind die Daten längsschnittlich verknüpfbar und erlauben dadurch Analysen von mehrjährigen Beobachtungszeiträumen. Außerdem können Abrechnungsdaten aus verschiedenen SGB-Kontexten, beispielsweise aus dem ambulanten oder stationären Bereich, sektorenübergreifend verwendet werden. Positiv ist zudem, dass insbesondere die abrechnungsessenziellen Daten vollständig und in geprüfter Form bei den Krankenkassen vorliegen. Auch im Projekt haben sich der große Datenumfang zu über einer Million Versicherten und die hohe Datenqualität als vorteilhaft erwiesen, um möglichst präzise Vorhersagemodelle entwickeln zu können.

- **Nachteile von Routinedaten:** Die Abrechnungsdaten aus den verschiedenen SGB-Leistungsbereichen liegen in einer relationalen Datenbank vor. Die Datenaufbereitung zu einem analysefähigen Datensatz gestaltete sich auch im Projekt sehr aufwändig und setzt nicht nur methodische Erfahrungen im Umgang mit Routinedaten, sondern auch Kenntnisse zur GKV-Abrechnungssystematik

und zu Spezifika der deutschen Gesundheitsversorgung voraus. Zudem benötigten die Datenaufbereitung und Umsetzung einzelner Ideen zur Optimierung der ML-Verfahren teilweise sehr viel personelle Ressourcen und Zeit. Dies erschwerte das explorative und nicht theoriegeleitete Analysieren der Daten, wie es im Bereich Big Data und KI verbreitet ist. Zudem enthielten die Analysedatensätze selbst bei einem theoriegeleiteten Vorgehen schnell eine Vielzahl potenziell relevanter Prädiktoren, weshalb einige ML-Methoden mit der zur Verfügung stehenden Hardware eine Rechenzeit von bis zu 14 Tagen benötigten. Zudem erwiesen sich die ML-Verfahren im Vergleich zur logistischen Regression weniger robust gegenüber sog. „unbalancierten Daten", wie es vor allem beim selten auftretenden Outcome Mortalität der Fall ist. Hier mussten die Daten bzw. die ML-Verfahren mit Hilfe spezieller Methoden (Up- und Downsampling, Verlustfunktion) zunächst angepasst werden, um brauchbare Vorhersagen zu erzielen.

10.3 Anwendbarkeit der Vorhersagemodelle in der Gesundheitsversorgung

- **Implementierung**: Das Vorläuferprojekt USER (Förderkennzeichen 01NVF18010) konnte bereits am Beispiel der logistischen Regression zeigen, dass diese grundsätzlich als Prognosemodell in die technischen Infrastrukturen von Krankenkassen implementiert und zur Berechnung von Vorhersagen für Versicherte genutzt werden kann. Während die logistische Regression auf einer vergleichsweise leicht zu implementierenden, mathematischen Gleichung basiert, benötigt es für die Anwendung der ML-Verfahren softwarespezifische „Container-Lösungen". Die Installation und Nutzung solcher Lösungen gilt es noch zu erproben. Zudem ist zu berücksichtigen, dass sich mit steigendem Datenumfang, der für die Prognoseberechnung erforderlich ist, auch die Komplexität der automatisierten Datenaufbereitungsprozesse auf Seiten der Krankenkassen erhöht.

- **Nutzung der Prognosen:** Im Rahmen des Vorläuferprojektes USER wurden die Prognosemodelle als neue Versorgungsform für das Entlassmanagement (EM) im Krankenhaus erprobt. Der Datenaustausch zwischen Krankenhäuser und Krankenkassen sowie die Prognoseberechnung erfolgten unmittelbar nach der stationären Aufnahme, so dass die Prognosen innerhalb von ein bis zwei Tagen den EM-Verantwortlichen als Entscheidungsunterstützung zur Verfügung gestellt werden konnten. Ein vergleichbares Anwendungssetting ist auch für den Einsatz der ML-gestützten Prognosen denkbar. Zudem lassen sich prinzipiell auch Modelle für andere Outcomes oder für einen anderen Bezugszeitraum als den Krankenhausaufenthalt berechnen, woraus sich zahlreiche, weitere Anwendungsmöglichkeiten ergeben. Ein aktuelles Beispiel wäre die Identifikation von Versicherten mit besonders hohen Gesundheitsrisiken gemäß § 25b SGB V. Grundsätzlich ist darauf hinzuweisen, dass der Einsatz solcher Vorhersageverfahren in der Regelversorgung aufgrund des Risikos der Fehlklassifikation immer auch aus ethischer und juristischer Perspektive zu bewerten ist.

- **Implikationen bei der Verwendung von GKV-Routinedaten:** Abrechnungsdaten liegen erst nach einem gewissen Zeitraum in geprüfter Form bei den Krankenkassen vor und können erst dann abgerufen werden. Dieser Datenverzug variiert zwischen den SGB-Leistungsbereichen und kann z. B. im Fall der ambulanten Abrechnungsdaten bis zu neun Monaten dauern. Diese eingeschränkte Datenaktualität gilt es sowohl für die Modellentwicklungsphase als auch bei der der späteren Berechnung von Prognosen „in Echtzeit" zu berücksichtigen. Darüber hinaus gelten die Routinedaten von Krankenkassen gemäß SGB als Sozialdaten und unterliegen somit hohen Auflagen. Für ihre Nutzung zu Forschungszwecken werden i.d.R. behördlichen Genehmigungen benötigt, deren Beantragung erfahrungsgemäß mehrere Monate dauert und die Arbeit in Forschungsprojekten mit einer begrenzten Laufzeit zuweilen erschwert. Zudem sieht die Gesetzgebung eine Einwilligung der Versicherten für die Verwendung ihrer Daten vor, was sich gerade bei den Versichertengruppen, die

besonders von derartigen KI-Lösungen profitieren könnten (z. B. multimorbide oder pflegebedürftige Menschen), als Hürde erweist. Hier bleibt abzuwarten, ob und inwieweit das Gesundheitsdatennutzungsgesetz (GDNG) künftig die Nutzbarkeit der GKV-Routinedaten für Verfahren der Künstlichen Intelligenz verbessert.

11 Anhang

11.1 Darstellungen von PR-Kurven mit Rückgriff auf Einzelbeobachtungen

Während die in Kapitel 3 dargestellte PR-Kurve in Abbildung 3-3 – wie stets auch alle ROC-Kurven – auf voraggregierten Ergebnissen zu Subgruppen mit jeweils unterschiedlichen vorhergesagten Risiken beruht, basieren viele Darstellungen von PR-Kurven offensichtlich auf Ergebnissen zu einzelnen Beobachtungen. Die Darstellung von Werten zu einzelnen Beobachtungen dürfte auch dadurch mitbedingt sein, dass KI-basierte Vorhersagen bei Beispieldarstellungen, aber auch in der Praxis, häufiger auf vergleichsweise wenigen Beobachtungen beruhen. Alle einzelnen Beobachtungen in den betrachteten Daten werden schrittweise berücksichtigt, woraufhin für die so neu entstandene Trennung der Daten auch ein Wertepaar aus Recall und Precision ermittelt und im Diagramm dargestellt wird. **Eine entsprechende Darstellung kann aufgrund folgender Beobachtungen grundsätzlich als problematisch angesehen werden.**

- Als Ergebnisse eines Prädiktionsmodells gehen in die Darstellung einer PR-Kurve (wie bei der ROC-Kurve) ausschließlich vorhergesagte Ereigniswahrscheinlichkeiten bzw. Risiken ein, die dazu verwendet werden, alle Beobachtungen zunächst absteigend nach diesen Risiken zu sortieren. Anschließend können für die PR-Kurve basierend auf den tatsächlich erfassten Ereignissen für unterschiedliche Splits der sortierten Beobachtungen dann Precision und Recall ermittelt werden.
- Typischerweise ist, besonders ausgeprägt bei Modellen mit Rückgriff auf wenige Merkmale im Sinne potenzieller Prädiktoren, davon auszugehen, dass nach den Modellvorhersagen bei jeweils mehreren Beobachtungen identische vorhergesagte Risiken resultieren. Innerhalb dieser Subgruppen liefert die Vorhersage dann keinerlei Kriterien für die Reihenfolge der Berücksichtigung von einzelnen Beobachtungen.
- Aus Variationen dieser Reihenfolge innerhalb von Subgruppen können jedoch sehr unterschiedliche Kurvendarstellungen resultieren, unter denen sich dann auch sehr unterschiedliche Flächen ermitteln ließen.

Den Sachverhalt verdeutlicht Abbildung 11-1. PR-Kurven mit Berücksichtigung von Ergebnissen zu einzelnen Beobachtungen, in der für Ergebnisse zu einer Modellvorhersagen Kurven mit drei unterschiedlichen Sortierungen von Beobachtungen innerhalb von Subgruppen präsentiert werden (vgl. auch Erläuterungen unter der Abbildung). Jeder Punkt in der Abbildung repräsentiert dabei das Ergebnis bei

Berücksichtigung einer weiteren Beobachtung der insgesamt 50 Beobachtungen. Ganz offensichtlich unterscheiden sich die Flächen unter diesen drei Kurven deutlich.

Ganz offensichtlich zeigen die drei Kurven zugleich auch allenfalls sehr bedingt die Precision der Vorhersage für einzelne Subgruppen. So besteht in diesem Beispiel die Gruppe mit den höchsten vorhergesagten Risiken aus 10 Personen, von denen 8 auch real betroffen waren. Für diese Gruppe lässt sich demnach unstrittig eine Precision von 0,8 angeben – 80 % dieser als „Hochrisikogruppe“ identifizierten Personen waren auch real betroffen. Die 8 Personen bilden dabei einen Anteil von 40 % der insgesamt 20 real betroffenen. Insofern sollte der Flächenanteil unter der Kurve bis zum Recall von 0,4 exakt 0,4 x 0,8 betragen, was hier für keine der drei Kurven zutrifft. Die Precision von 0,8 lässt sich ganz offensichtlich aus keiner der drei Kurven intuitiv ablesen.

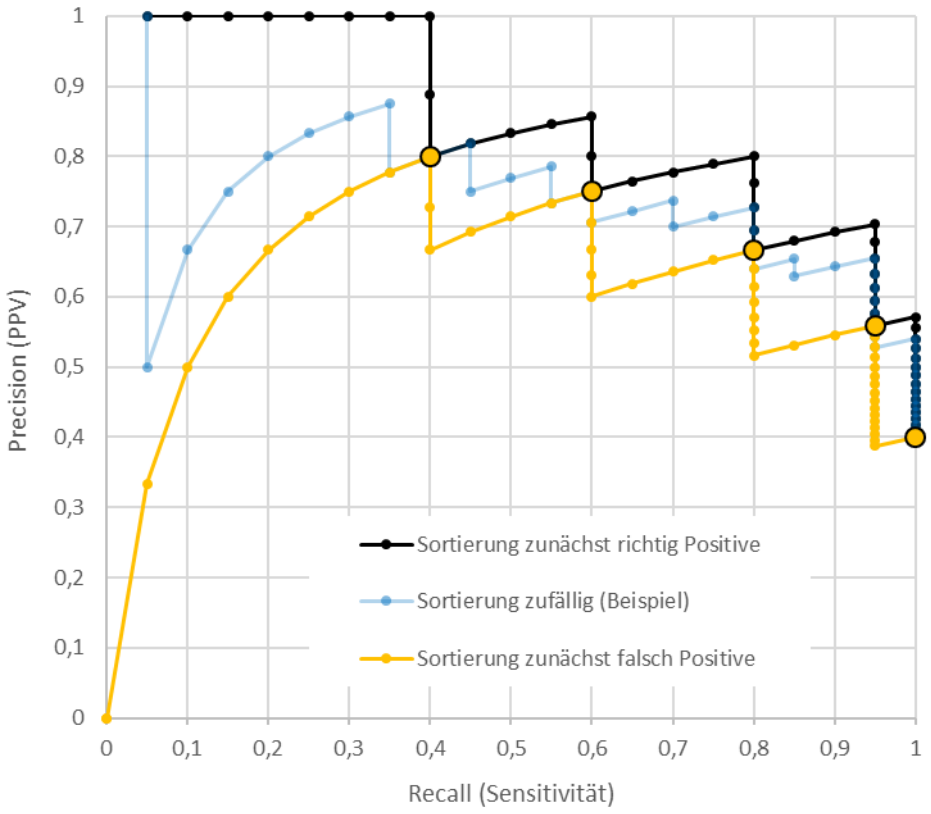

Abbildung 11-1. PR-Kurven mit Berücksichtigung von Ergebnissen zu einzelnen Beobachtungen

Beispiel mit drei Kurvendarstellungen zu einem Modellergebnis bei 50 Beobachtungen verteilt auf 5 Subgruppen mit unterschiedlichen vorhergesagten Risiken bei Variation der Sortierung innerhalb von Subgruppen: a) zunächst Berücksichtigung der real Betroffenen, b) Beispiel einer zufälligen Sortierung sowie c) zunächst systematische Berücksichtigung der real nicht Betroffenen; Hervorhebung von Datenpunkten, die allen drei Varianten gemein sind (es handelt sich dabei jeweils um Ergebnisse zur letztberücksichtigten Beobachtung mit einem bestimmten vorhergesagten Risiko)

Nach Sichtung von Darstellungen im Internet findet für Darstellungen der PR-Kurve basierend auf Einzelbeobachtungen oftmals Variante a) eine Anwendung. Nur damit ist nachvollziehbar, weshalb es das in empirischen Analysen eher unwahrscheinliche Ergebnis einer Precision = 1 in den Darstellungen relativ häufig gibt – entsprechende Kurven stellen (zumindest weit überwiegend) die eigentliche Precision der Modellvorhersage nicht korrekt dar und sind insofern als „Precision Recall-Kurven“ irreführend. **Aus diesen Beobachtungen folgt, dass die Darstellung einer Kurve basierend auf Ergebnissen**

zu Precision und Recall mit Differenzierung von Ergebnissen nach Einzelbeobachtungen eigentlich eher nicht als „Precision Recall-Kurve“ bezeichnet werden sollte.

Darstellungen von PR-Kurven mit Rückgriff auf aggregierte Ergebnisse zu Subgruppen mit unterscheidbaren vorhergesagten Risiken

Aus den vorausgehenden Erläuterungen folgt, dass (auch) die Darstellung einer PR-Kurve (wie Darstellungen zur ROC-Kurve) vorzugsweise auf aggregierten Daten basieren sollte, sofern hieraus die Precision bei Erreichen eines bestimmten Recall ablesbar sein und damit eine inhaltlich bedeutsame Fläche dargestellt werden soll. Wie eine derartige Kurve verlaufen sollte, deuten die Ergebnisse in Abbildung 11-1 an: Offensichtlich existieren Datenpunkte, die, unabhängig von der Sortierung der Beobachtungen innerhalb von Subgruppen mit identisch vorhergesagten Risiken, in allen drei Varianten identisch sind – nämlich die Punkte, die nach Berücksichtigung jeweils aller Beobachtungen mit einem bestimmten vorhergesagten Risiko resultieren. Diese Punkte stellen die Precision für eine begrenzte Auswahl der Recall-Werte in der Kurve bereits korrekt dar. Genau diese Auswahl an Punkten kann auch aus den aggregierten Ergebnissen abgelesen werden, welche jeweils die Ergebnisse zu allen Beobachtungen mit einem identisch vorhergesagten Risiko zusammenfassen. Damit verbleibt lediglich die Beantwortung der Frage, wie die Kurve zwischen diesen Datenpunkten verlaufen sollte.

Für den ersten Abschnitt dieser Beispielkurve wurde die Frage bereits im vorausgehenden Text beantwortet. Für die im ersten Kurvenabschnitt von Abbildung 11-1 dargestellten Ergebnisse zur „Hochrisikogruppe“ erreicht die Modellvorhersage im Beispiel eine Precision von 0,8. Dieser Sachverhalt wird korrekt durch eine waagerechte Strecke vom Punkt (x = 0; y = 0,8) bis zum Punkt (x = 0,4; y = 0,8) dargestellt.

Für den weiteren Kurvenverlauf erscheint es naheliegend, die nachfolgenden Punkte durch Geraden zu verbinden, wie dies als Vorgehensweise bei der ROC-Kurve erfolgt und dabei auch korrekt ist. Entsprechend wurde der Verlauf in Abbildung 3-3 dargestellt. Allerdings ist dies bei einer PR-Kurve nur annähernd und nicht vollständig korrekt. Abweichungen resultieren daraus, dass zwischen Recall und Precision kein einfacher linearer Zusammenhang wie zwischen Falsch-Positiv-Rate und Recall bei der der ROC-Kurve besteht, was mit den nachfolgenden theoretischen Überlegungen zur Berechnung der Fläche unter der PR-Kurve verdeutlicht werden soll.

Berechnung der Fläche unter der PR-Kurve

Will man Abhängigkeiten der resultierenden Darstellung einer PR-Kurve von willkürlichen Sortierungen der Einzelbeobachtungen innerhalb der Subgruppen mit identischen vorhergesagten Risiken wie in Abbildung 11-1 vermeiden, lässt sich dies dadurch bewerkstelligen, dass schrittweise Beobachtungen bei Berechnungen von Punkten im Verlauf berücksichtigt werden, denen anstelle der tatsächlichen Ereigniswerte 0 = nicht betroffen oder 1 = betroffen jeweils einheitlich die durchschnittlich beobachtete Betroffenenrate in der jeweiligen Subpopulation zugeordnet wird (was gemäß Summentreue des Mittelwertes (MW) bei Summation zum selben Ergebnis wie die Addition der Ereigniswerte in der Subgruppe führt). Erreicht wird durch die Berücksichtigung von durchschnittlichen Risiken bzw. MW die Darstellung eines „idealisierten“ Kurvenverlaufs, der entsprechend auch bei in Subgruppen rein zufällig sortierten Beobachtungen erwartbar wäre, wenn die Größe einer „idealerweise“ betrachteten Population gegen unendlich streben würde (oder wenn eine gegen unendlich strebende Anzahl an Ergebnissen basierend auf zufällig angeordneten Beobachtungen in Subgruppen wie in Abbildung 11-1**Fehler! Verweisquelle konnte nicht gefunden werden.** gemittelt würde), womit zufallsbedingte Variationen der Reihenfolgen von Beobachtungen, anders als in dem einen Beispiel der zufällig sortierten Beobachtungen in Abbildung 11-1 auch ohne eine Berücksichtigung von Mittelwerten bei der Darstellung nicht mehr ins Gewicht fallen würden. Die Berücksichtigung von MW bietet zudem den Vorteil, dass Darstellungspunkte für eine Abbildung, auch unabhängig von der tatsächlichen Beobachtungszahl in der Untersuchungspopulation, in beliebig erhöhter oder auch reduzierter Zahl errechnet werden können. Ergebnisse zu einem sehr einfachen, fiktiven Beispiel zeigt Abbildung 11-2.

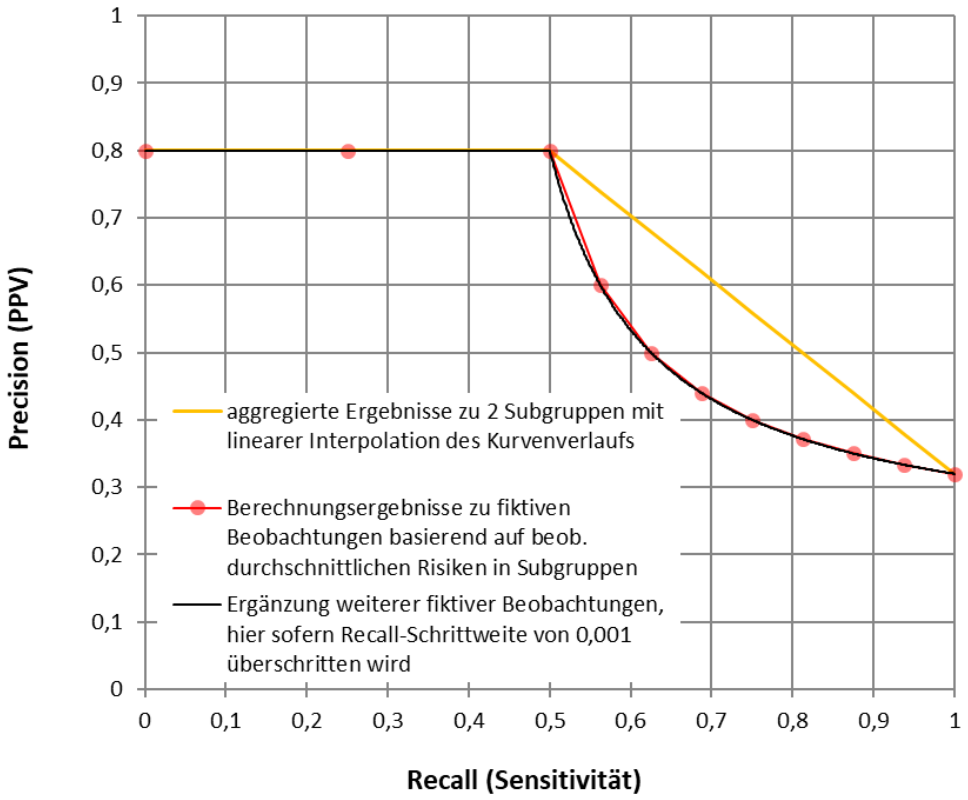

Abbildung 11-2. PR-Kurven zu einem Datenbeispiel mit Rückgriff auf aggregierte Ergebnisse sowie mit Berücksichtigung von Mittelwerten (MW) zu einzelnen Beobachtungen

(Fiktives Beispiel, bei dem Vorhersagen lediglich zwei Subgruppen differenzieren und von n = 20 Beobachtungen der Hochrisikogruppe real 16 (entsprechend einem Anteil von 80 %) und von n = 80 Personen der Niedrigrisikogruppe real gleichfalls 16 (entsprechend einem Anteil von 20 %) betroffen sind, womit sich für die insgesamt n = 100 Beobachtungen ein Betroffenenanteil von 32 % ergibt.)

Die erste (orange und teils verdeckte) Kurve verdeutlicht eine Darstellung basierend ausschließlich auf aggregierten Ergebnissen zu den beiden Subgruppen, wobei zwischen den hier in aggregierten Ergebnissen aufgezeigten Punkten der Kurve zu Subgruppe 1 und 2 (bei einem Recall von 0,4 bzw. 1,0) vereinfacht und approximativ linear interpoliert wird.

Die zweite (rote) Kurve zeigt gleichfalls eine Darstellung ausschließlich basierend auf aggregierten Ergebnissen, wobei hier jedoch (fiktive, idealtypische) Ergebnisse zu jeder 10. Beobachtung mit Rückgriff auf die durchschnittlich real beobachteten Risiken in den jeweiligen Subgruppen errechnet und dargestellt sind. Dabei weichen die Darstellungen der beiden Kurven in diesem Beispiel deutlich voneinander ab. Offensichtlich führt in dem Beispiel eine konsekutive Berücksichtigung bereits von relativ wenigen Beobachtungen aus der zweiten Gruppe (mit durchschnittlich erheblich geringeren Risiken) zu einem deutlich stärkeren Rückgang der Precision, als dies durch die lineare Interpolation in der ersten Kurve zum Ausdruck kommt – die lineare Interpolation überschätzt in diesem Fall die Precision also merklich. Eine weitgehend idealtypische Kurve resultiert durch die Interpolation weiterer Punkte im Kurvenverlauf, wobei die bereits im rot gefärbtem Kurvenverlauf dargestellten Punkte erhalten bleiben (vgl. schwarze Linie). Die Fläche unter dieser idealtypischen Kurve entspricht dann dem korrekt ermittelten Kennwert, wie er sich z. B. auch mit der entsprechenden Python-Funktion aus dem Package scikit-learn (precision_recall_curve) oder in R mit dem package PRROC und der Funktion pr.curve errechnen lässt.

11.2 Soft- und Hardware im Projekt KI-THRUST

Für die Lernprozesse bei KI-Verfahren müssen Daten in hochskalierbaren Rechenabläufen verarbeitet werden. Um diese Berechnungen effizient durchführen zu können, ist eine geeignete technische Infrastruktur erforderlich, die sowohl auf leistungsfähiger Hardware als auch auf spezialisierter Software basiert. Insbesondere die angestrebten Verfahren des maschinellen Lernens benötigen eine hohe Rechenleistung, welche durch die Nutzung von GPUs (Grafikkarten) und leistungsstarken Prozessoren gewährleistet wird.

Der Rechenaufwand lässt sich vor allem anhand der Anzahl der zu berücksichtigenden Merkmale (Parameter) und der Komplexität der jeweiligen Modelle (KI-Verfahren) abschätzen. Bei der Verarbeitung von Sozialdaten ist davon auszugehen, dass eine Vielzahl von Parametern einbezogen wird, die in tiefen künstlichen Neuronalen Netzen verarbeitet werden sollen.

Die eingesetzten Softwarekomponenten gliedern sich wie folgt:

- Programmiersprache Python mit der Bibliothek Keras (Deep Learning)
- Entwicklungsumgebung Jupyter Notebook
- Backend TensorFlow mit einer CUDA-Schnittstelle für Parallelberechnungen auf mehreren GPUs

Die Hardwarekomponenten sind in folgender Tabelle 11-1 aufgelistet und beschrieben.

Tabelle 11-1. KI-Workstation-Spezifikationstabelle

Komponente	Spezifikation	Beschreibung
Prozessor (CPU)	AMD Threadripper 3970X	32 Kerne, 64 Threads, 3,7 GHz Basis, 4,5 GHz Boost, 144 MB Cache, 280W TDP
Arbeitsspeicher (RAM)	8x 16GB HyperX Fury DDR4 3200 MHz (128GB gesamt)	Hochgeschwindigkeits-DDR4-Speicher mit 3200 MHz
Grafikkarte (GPU)	PNY Nvidia RTX A6000 (48GB VRAM)	High-End-Workstation-GPU mit 48GB GDDR6-Speicher
Mainboard	Gigabyte Aorus Master TRX40	TRX40-Chipsatz-Mainboard für Threadripper-CPUs
Netzteil (PSU)	Be Quiet! Dark Power 12 1200W	80 PLUS Titanium Effizienz, vollmodular
Primärer Speicher (SSD)	2x 2TB WD Black SN770 (4TB gesamt)	Hochgeschwindigkeits-NVMe-SSDs mit PCIe 4.0
Netzwerkspeicher	QNAP TS-832PXU-RP	8-Bay-Rackmount-NAS
NAS-Festplatten		4x 2TB Seagate IronWolf (8TB gesamt)

Konzept für eine projektspezifische IT-Infrastruktur

1. Ausgangslage

Für das Innovationsfondsprojekt Projekt „KI-THRUST“ (Förderkennzeichen 01VSF20014) ist die Anwendung komplexer Verfahren der Künstlichen Intelligenz (KI) auf der Grundlage von umfangreichen GKV-Routinedaten vorgesehen. Hierfür wird für das Projekt eine leistungsfähige IT-Infrastruktur benötigt, die sowohl den hohen Anforderungen im Bereich des maschinellen Lernens gerecht wird als auch eine Integration aktueller Technologien ermöglicht. Insbesondere wird eine Hard- und Softwarelösung angestrebt, die ausreichend Leistungskapazität für das Training und die Ausführung von rechenintensiven KI-Modellen bietet. Hinsichtlich der Software besteht die Vorgabe, dass die Infrastruktur vollständig kompatibel mit einer plattformunabhängigen Open-Source-Programmbibliothek betrieben werden soll. Hierbei entschied sich das Projektkonsortium frühzeitig für die Nutzung von „TensorFlow“, einem Framework für datenstromorientierte Programmierung, welches vor allem im Bereich des maschinellen Lernens breite Anwendung findet und eine effiziente Entwicklung ermöglicht.

2. Spezifikation der Hardware

Bereits vor dem Projektstart im Juli 2021 erfolgten im Rahmen einer Konzeptionierungsphase regelmäßige Abstimmungen zwischen der IT-Abteilung des aQua-Instituts, die verantwortlich für die Beschaffung und den Aufbau der IT-Infrastruktur gewesen ist, und dem Konsortialpartner der Universitätsmedizin Göttingen (UMG), die die geplanten KI-Analysen durchführt. Erste Diskussionen drehten sich um die Frage, ob für das Projekt Hardware aus dem Consumer- oder Enterprise-Segment eingesetzt werden sollte (siehe Tabelle 1). Dazu wurden initial verschiedene Konfigurationsvarianten erstellt, die sich sowohl hinsichtlich Rechenleistung und Speicherarchitektur als auch in Bezug auf Größe, Energiebedarf und Gesamtkosten unterschieden. Zudem wurden Schätzungen zu voraussichtlichen Betriebskosten, insbesondere zum Stromverbrauch, vorgenommen.

Tabelle 1: Konfigurationsvarianten der Hardware

Komponente	Enterprise	Consumer	Final
Prozessor (CPU)	Intel® Xeon® W-3323 Prozessor	Intel Core i7-12700K	AMD Threadripper 3970X
Arbeitsspeicher (RAM)	8x 32GB Kingston X-Tream DDR4 2666 (256GB gesamt)	64 GB	8x 16GB HyperX Fury DDR4 3200 MHz (128GB gesamt)
Grafikkarte (GPU)	2x PNY Nvidia RTX A6000 (48GB VRAM)	RTX 2080 TI	PNY Nvidia RTX A6000 (48GB VRAM)
Mainboard	ASUS WS X299 SAGE	Gigabyte Z490	Gigabyte Aorus Master TRX40
Netzteil (PSU)	Be Quiet! Dark Power 12 1200W	Be Quiet! Dark Power 12 1200W	Be Quiet! Dark Power 12 1200W
Primärer Speicher (SSD)	2x 4TB NVME Samsung Pro 890 (8TB gesamt)	4x 2TB WD Black SN770 (8TB gesamt)	2x 2TB WD Black SN770 (4TB gesamt)
Netzwerkspeicher	QNAP TS-832PXU-RP	QNAP TS-832PXU-RP	QNAP TS-832PXU-RP
NAS-Festplatten	4x 2TB Seagate IronWolf (8TB gesamt)	4x 2TB Seagate IronWolf (8TB gesamt)	4x 2TB Seagate IronWolf (8TB gesamt)

Während die Enterprise-Komponenten deutliche Vorteile im Hinblick auf Performance und Stabilität bietet, sind sie zugleich mit deutlich höheren Kosten verbunden. Nach Kosten-Nutzen-Abwägung fiel die Entscheidung zunächst zugunsten der Komponenten der Consumer-Variante, die als ausreichend skalierbar eingeschätzt wurde und somit den Projektanforderungen – trotz geringerer Leistung im Vergleich zur Enterprise-Variante – genügen dürfte.

Grafikkarte

Bei der Auswahl der Grafikkarte(n) standen zunächst zwei Nvidia RTX 2080 Ti zur Diskussion. Letztlich fiel die Wahl allerdings im Hinblick auf Skalierbarkeit und Speicheranforderungen sowie die Lizensierungsoptionen auf die leistungsstärkere Nvidia RTX A6000 aus dem Enterprise-Bereich. Ausschlaggebend waren dabei der größere Grafikspeicher, die höhere Anzahl an Recheneinheiten sowie die Tatsache, dass Nvidia die professionelle Nutzung der A6000 zu diesem Zeitpunkt lizenzrechtlich ausdrücklich erlaubte.

Prozessor

Neben der Auswahl der Grafikkarte spielten auch die übrigen Hardwarekomponenten eine wichtige Rolle. Beim Prozessor (CPU) lag der Fokus auf einer möglichst hohen Anzahl an Rechenkernen, um eine effiziente Parallelverarbeitung zu ermöglichen. Die Wahl fiel daher auf den AMD Ryzen Threadripper 3970x, der mit 32 physischen Kernen (mit 64 Threads) sowie einem Takt von je 3,7 bis 4,5 GHz eine hohe Rechenleistung bereitstellt. Trotz der leistungsstarken Architektur liegt die CPU preislich deutlich unter vergleichbaren Enterprise-Modellen. Die Threadripper-Plattform gehört zum oberen Segment des professionellen Consumer-Markts und wird beispielsweise für den professionellen Videoschnitt verwendet. So ist die CPU hervorragend geeignet für intensive Workloads einschließlich solcher auf Basis von „TensorFlow“. Damit erfüllt sie die Anforderungen des Projekts vollumfänglich im Hinblick auf Performance, Skalierbarkeit und Wirtschaftlichkeit.

Mainboard

Als Mainboard wurde sich für das Gigabyte TRX40 entschieden, das nicht nur volle Kompatibilität zur CPU gewährleistet, sondern auch sämtliche Anforderungen an Schnittstellen und Systemstabilität erfüllt. Wichtig war bei der Auswahl auch die Möglichkeit, die elektrischen Komponenten auf dem Board effizient zu kühlen. Das TRX40 bietet hierfür insgesamt sieben 4-Pin-PWM-Anschlüsse auf der Hauptplatine, die eine flexible Ansteuerung von Lüftern und Wasserpumpen ermöglichen. Dies schafft die Grundlage für ein leistungsfähiges und stabiles System, was einen zentralen Faktor im Hinblick auf rechenintensive Workloads ausmacht.

Arbeitsspeicher

Für den Arbeitsspeicher wurden 128 GB DDR4 mit EEC-Funktionalität (Error-Correcting Code) verbaut. Die ECC-Unterstützung sorgt für eine automatische Fehlerkorrektur im laufenden Betrieb und erhöht damit die Systemsicherheit bei großen Datenmengen. Die gewählte Speicherkapazität ist für das Projekt ausreichend, lässt sich jedoch bei Bedarf problemlos erweitern.

Kühlung

Damit die Hardware entsprechend gekühlt werden kann, kommt ein Wasserkühlungskompressor mit einem Fassungsvermögen von 9 Litern zum Einsatz. Dieser ist darauf ausgelegt, sowohl die CPU als auch die GPU auch unter hoher Last zuverlässig zwischen 20° und 30° zu halten. Durch diesen Kühlprozess wird die Hardware vor einer Überhitzung geschützt und die Langlebigkeit sowie der energieeffiziente Betrieb der einzelnen Komponenten dauerhaft unterstützt.

Das restliche System ist zum Großteil aus Consumer-Hardware aufgebaut. Damit konnten die Kosten enorm gesenkt werden, ohne dass auf Stabilität und Effizienz verzichtet werden musste.

Besonderheiten zu Projektbeginn

Unmittelbar nach dem Projektbeginn im Juli 2021 wurde die Hardwarestrategie erneut evaluiert. Grund hierfür war die anhaltende globale Chipknappheit, die sowohl Verfügbarkeiten als auch Preise vieler Komponenten stark beeinflusste. Die IT-Abteilung nahm daher eine aktualisierte Marktanalyse vor, wo Skalierbarkeit und Kosten ausschlaggebend für die endgültige Beschaffung blieben. An dem ursprünglichen Beschaffungsplan wurde nach dieser Evaluierung festgehalten.

3. Spezifikation der Software

Betriebssystem

Auf Seiten der Software wurde sich für die Linux-Distribution Ubuntu (20.04 LTS) entschieden. Dadurch können aktuelle Softwarepakete aus offiziellen Paketquellen bezogen werden. Die LTS-Version (Long Time Service) bietet zudem den Vorteil, dass bis 2025 Wartungsupdates und bis 2030 Sicherheitsupdates erfolgen. Damit ist innerhalb der Projektlaufzeit kein Upgrade des Betriebssystems erforderlich.

Arbeitsumgebung

Um Probleme mit Softwareabhängigkeiten zu vermeiden, wurde die Nutzung von TensorFlow in Kombination mit der Software „Docker“ gewählt, die Container-Virtualisierung von Anwendungen ermöglicht. Anwendungen können dabei inklusive aller Abhängigkeiten in einem Image gebündelt, transportiert und installiert werden. Zunächst stellt NVIDIA ein Docker-Toolkit bereit, mit dem die Leistung der Grafikkarte direkt den Docker-Containern zur Verfügung gestellt werden kann. Das Projekt läuft somit in einer standardisierten Umgebung.

Docker wurde gemäß der empfohlenen Sicherheitsrichtlinie („Best Practice“) im sogenannten „Non-Root-Modus“ konfiguriert. Daraufhin wurde eine dezidierte Benutzergruppe „docker“ angelegt. Externe Nutzer der UMG (Universität Göttingen) wurden dieser Gruppe zugewiesen, sodass sie mit Docker arbeiten können, jedoch über keine root-Rechte auf dem System verfügen.

Zur konkreten Umsetzung wurde für das TensorFlow-Workspace, inkl. Ereignissen, Skripten und weitere Ressourcen, ein separates Verzeichnis angelegt, auf das die Gruppe „docker“ Zugriff hat. Darin befindet sich auch die „docker-compose.yml“, um TensorFlow mit einer GPU-Unterstützung als Docker Container zu starten.

Treiber

Zudem musste der vorinstallierte Standardtreiber für Nvidia Grafikkarten „nouveau“ deaktiviert werden, da dieser mit dem nicht quelloffenen Nvidia-Treiber inkompatibel ist. Erst nach der Deaktivierung ist es somit möglich, den offiziellen „nonfree“-Nvidia-Treiber zu installieren.

Anschließend wurde der Nvidia-Treiber für Linux über das offizielle run-Installationsskript installiert und Docker (inklusive Docker Compose) in der Community Edition sowie das Paket „nvidia-docker2“ aus den Ubuntu-Paketquellen eingerichtet. Dies ermöglicht es, Nvidia-GPUs innerhalb von Docker-Containern zu nutzen.

Sicherheit & Zugriffsrechte

Das System wurde zuerst an die unternehmensinternen IT-Standards angepasst, um Sicherheits- und Verwaltungsrichtlinien einzuhalten. Dazu wurde der SSH-Port auf 2222 gestellt, um automatisierte Angriffe auf den Standardport zu erschweren. Zudem wurde der direkte root-Zugang deaktiviert, um

potenzielle Sicherheitsrisiken durch unkontrollierten Zugriff zu minimieren. Stattdessen erfolgt der Zugriff über reguläre Benutzerkonten mit klar zugewiesenen Rechten.

Es wurde eine Firewall mit UFW (Uncomplicated Firewall) konfiguriert, die Uhrzeit wurde über NTP (Network Time Protocol) mit offiziellen Zeitservern synchronisiert und das Monitoring-System Zabbix wurde eingerichtet. Letzteres dient zur kontinuierlichen Überwachung des Zustands und der Leistung des Systems und beinhaltet ein automatisches Benachrichtigungssystem bei Auffälligkeiten.

Testphase

Die Stabilität des hier vorgestellten Systems wurde nach Aufbau und Einrichtung über eine Dauer von drei Monaten durch mehrere Belastungstests in Form von exemplarischen Rechenaufgaben sichergestellt. Hier zeigte sich auch, dass die Performance die geforderten Ansprüche vollends erfüllt. Erst nach diesen drei Monaten zeigte sich, dass einige Pakete innerhalb des Test-Containers ein Update benötigten, was über ein Redeploy des Test-Containers innerhalb weniger Sekunden erfolgen konnte.

aQua
ZUKUNFT DURCH QUALITÄT

Stand: 23.05.2022

Datensatzbeschreibung zum Projekt KI-THRUST

Allgemeine Erläuterungen

Die folgende Datensatzbeschreibung enthält alle notwendigen Daten, die für das vom Innovationsausschuss beim Gemeinsamen Bundesausschuss (G-BA) geförderte Projekt KI-THRUST (Förderkennzeichen 01VSF20014) benötigt werden.

1. Pseudonymisierung

Die Pseudonyme zu den Versicherten und zu den Leistungserbringern müssen über alle Datenjahre konsistent sein. Jegliche Kennzeichnungen von Personen oder Einrichtungen (z.B. Krankenhäuser, Betriebsstätten) erfolgen durchgängig in einer Form, die es dem aQua-Institut nicht erlaubt, Rückschlüsse auf identifizierbare Personen oder Einrichtungen zu ziehen. Dabei sollen beispielsweise Betriebsstättennummern jeweils durch eine kassenintern eindeutig zugeordnete 14-stellige Ziffern- oder Zeichenfolge ersetzt werden, welche keinerlei Rückschlüsse auf die ursprüngliche Betriebsstättennummer erlaubt.

2. Selektionskriterien

Maßgeblich für die Selektion der Leistungsdaten sind alle Versicherten der teilnehmenden Krankenkassen, die in den Jahren 2015 bis einschließlich 2020 eine Entlassung aus einer stationären Krankenhausbehandlung gehabt haben. Berücksichtigt werden nur geprüfte und abgeschlossene Fälle.

Es gibt keine weiteren Ein- oder Ausschlusskriterien.

3. Datentabellen

Für die selektierten Patienten sollen Daten der nachfolgend beschriebenen Tabellen bereitgestellt werden. Die den Namen der Tabellen vorangestellten Ziffern beziehen sich auf die gesetzliche Grundlage des SGB. Die angeforderten Datenfelder sind angelehnt an die technischen Anlagen des Datenaustauschverfahrens und deren Schlüsselverzeichnisse (wenn solche existieren).

4. Technische Hinweise zur Datenbereitstellung und Datenübermittlung

Alle nachfolgend spezifizierten Datentabellen sollen als einfache Textdateien bereitgestellt werden. Dabei soll eine UTF-8-Zeichenkodierung verwendet werden. Als Trennzeichen für bereitgestellte Datenfelder/Werte innerhalb von einzelnen Datenzeilen (im Sinne des Kommas bei CSV-Dateien in einem engeren Sinne) soll ein Semikolon (;) verwendet werden. Die Benennung der einzelnen Textdateien bzw. Datentabellen (z.B. Stammdaten.csv) sowie die Formate und die Benennungen von Variablen sind durch die nachfolgenden Spezifikationen vorgegeben.

Für eine gesicherte und datenschutzkonforme Übermittlung der selektierten und pseudonymisierten Daten wird seitens des aQua-Instituts ein gesicherter SFTP-Zugang auf ein individuelles Laufwerksverzeichnis im aQua-Institut eingerichtet werden. Dazu muss von der datenbereitstellenden Institution jeweils eine feste IP-Adresse angegeben werden, über welche die Institution die Daten übermitteln will. Die Übermittlung der Daten kann dann nach der Bereitstellung von Zugangsdaten sowie der persönlichen Mitteilung eines Passwortes per Telefon durch das aQua-Institut durch die bereitstellende Institution/Krankenkasse über die festgelegte IP-Adresse mittels eines üblichen SFTP-Clients (z.B. FileZilla) erfolgen.

Tabelle 1: Stammdaten

Name	ID	Type	Comments
VERSID	1	alphanumerisch	Versichertenpseudonym
Geburtsjahr	2	numerisch (3)	Geburtsjahr (JJJJ)
TOD_Datum	3	DATUM (8)	Todesdatum (JJJJMMTT) sofern zutreffend
GESCHL	4	alphanumerisch (1)	Geschlecht (m=männlich; w=weiblich; x=inter/divers/unbekannt)
PLZ_3_stellig	7	alphanumerisch	3-stellige aktuelle PLZ des Wohnorts des Versicherten

Tabelle 2: Versicherungszeiten

Name	ID	Type	Comments
VERSID	1	alphanumerisch	Versichertenpseudonym
VERS_Start	2	DATUM (8)	Beginn der Versicherungszeit (JJJJMMTT)
VERS_Stop	3	DATUM (8)	Ende der Versicherungszeit (JJJJMMTT)

Anmerkung: Die Versicherungszeiten sind bei Unterbrechungen ggf. in mehreren Zeilen pro Versicherten zu übermitteln. Je nach Datenlieferung sollen alle Versicherungsintervalle mit Relevanz für die zu liefernden Datenjahre übermittelt werden (siehe Tabelle, Seite 1). So umfasst beispielsweise die erste Datenlieferung alle Versicherungsintervalle für die Jahre 2015 bis 2017, insofern VERS_start <= 31.12.2017 UND VERS_stop >= 01.01.2015.

Tabelle 3: 295 AMB_FALL

Name	ID	Type	Comments
JAHR	1	numerisch (4)	Berichtsjahr
QUARTAL	2	numerisch (1)	Berichtsquartal
VERSID	3	alphanumerisch	Versichertenpseudonym
VERTRAGSKENNZEICHEN	4	alphanumerisch (1)	K=kollektivvertraglich; S=selektivvertraglich
BSNR	5	alphanumerisch	Betriebsstättennummer, pseudonymisiert
FALLNR	6	numerisch	Behandlungsfallnummer
BEH_VON	7	DATUM (8)	Beginn des Abrechnungsfalles (JJJJMMTT)
BEH_BIS	8	DATUM (8)	Ende des Abrechnungsfalles (JJJJMMTT)
INANSPR	9	alphanumerisch (1)	Art der Inanspruchnahme "O" Originalschein (Default) "V" Vertreterschein "N" Notfallschein "Z" Zielauftrag "K" Konsiliarauftrag "M" Mit-/ Weiterbehandlung

Tabelle 4: 295 AMB_ICD

Name	ID	Type	Comments
VERSID	1	alphanumerisch	Versichertenpseudonym

FALLNR	2	numerisch	Behandlungsfallnummer
ICD_LOK	3	alphanumerisch (1)	Seitigkeit der Diagnose (L=Links, R=Rechts, B=Beidseitig)
ICD_QUAL	4	alphanumerisch (1)	Diagnosekennzeichen (A=Ausschluss, V=Verdacht auf, Z=Zustand nach, G=Gesichert)
ICD	5	alphanumerisch (6)	Diagnose ohne Seitigkeit und ohne Sonderkennzeichen

Tabelle 5: 295 AMB_EBM

Name	ID	Type	Comments
VERSID	1	alphanumerisch	Versichertenpseudonym
LANR	2	alphanumerisch	Lebenslange Arztnummer, pseudonymisiert (falls vorhanden)
FALLNR	3	numerisch	Behandlungsfallnummer
GONR	4	alphanumerisch (7)	Gebührenordnungsnummer
BEHANDLUNGSDATUM	5	DATUM (8)	Format: JJJJMMTT

Tabelle 6: 295 AMB_OPS

Name	ID	Type	Comments
VERSID	1	alphanumerisch	Versichertenpseudonym
FALLNR	2	numerisch	Behandlungsfallnummer
OPS	3	alphanumerisch (1)	OPS-Schlüssel in der jeweils gültigen Fassung des DIMDI
OPS_LOK	4	alphanumerisch (1)	Diagnosekennzeichen (A=Ausschluss, V=Verdacht auf, Z=Zustand nach, G=Gesichert)

Tabelle 7: 300 VO

Name	ID	Type	Comments
VERSID	1	alphanumerisch	Versichertenpseudonym
LANR	2	alphanumerisch	Lebenslange Arztnummer, pseudonymisiert (falls vorhanden)
BSNR	3	alphanumerisch	Betriebsstättennummer des verordnenden Arztes, pseudonymisiert
PZN	4	numerisch	Pharmazentralnummer (für Hilfsmittel gelten die Sonderkennzeichenregelungen nach technischen Anlagen für den Datenaustausch gemäß aktuell gültiger Fassung)
ATC_CODE	5	alphanumerisch (7)	Amtlicher ATC-Code
DDD	6	alphanumerisch (9)	Daily Defined Dose
MULT	7	numerisch	Multiplikationsfaktor
DATUM_VO	8	DATUM (8)	Einlösedatum Format: JJJJMMTT
FACHGRUPPE	9	alphanumerisch (2)	Fachgruppe des Arztes (bei mehrfachen Fachgruppen je Arzt den Minimum-Wert)

Tabelle 8: 301 KH_AMB_FALL

Name	ID	Type	Comments
VERSID	1	alphanumerisch	Versichertenpseudonym
IK	2	alphanumerisch	IK des Krankenhauses, pseudonymisiert
ID	3	numerisch	Fall-ID
ZUGANG_DATUM	4	DATUM (8)	Aufnahmetag/Tag des Zugangs (JJJJMMTT)
ENTLASS_DATUM	5	DATUM (8)	Letzter Tag der Behandlung JJJJMMTT (entspricht bei ambulantem Operieren dem höchsten Behandlungstag

Tabelle 9: 301 KH_AMB_ICD

Name	ID	Type	Comments
VERSID	1	alphanumerisch	Versichertenpseudonym
IK	2	alphanumerisch	IK des Krankenhauses, pseudonymisiert
ID	3	numerisch	Fall-ID
ICD_DIAG	4	alphanumerisch (6)	Diagnose ohne Seitigkeit und ohne Sonderkennzeichen
ICD_DIAG_LOK	5	alphanumerisch (1)	Lokalisation / Diagnose: "L" = links "R“ = rechts "B" = beide
ICD_ART	6	numerisch (1)	1 = Behandlungsdiagnose, 2 = bei § 116b (neu): Diagnose der Überweisung innerhalb der ASV
ICD_SEK_DIAG	7	alphanumerisch (6)	Sekundärdiagnose
ICD_SEK_DIAG_LOK	8	alphanumerisch (1)	Sekundärdiagnose Lokalisation / Diagnose: "L" = links "R“ = rechts "B" = beide

Tabelle 10: 301 KH_AMB_OPS

Name	ID	Type	Comments
VERSID	1	alphanumerisch	Versichertenpseudonym
ID	2	numerisch	Fall-ID
OPS	3	alphanumerisch (8)	Amtlicher OP-Schlüssel nach jeweils gültiger Fassung
OPS_LOK	4	alphanumerisch (1)	Lokalisation / Diagnose: "L" = links "R“ = rechts "B" = beide

Tabelle 11: 301 KH_AMB_EBM

Name	ID	Type	Comments
VERSID	1	alphanumerisch	Versichertenpseudonym
ID	2	numerisch	Fall-ID
Entgeltschluessel	3	alphanumerisch (8)	KH - Entgeltschlüssel

Tabelle 12: 301 KH_FALL

Name	ID	Type	Comments
VERSID	1	alphanumerisch	Versichertenpseudonym
IK	2	alphanumerisch	IK des Krankenhauses, pseudonymisiert
ID	3	alphanumerisch	Fall-ID
AUFN_DATUM	4	DATUM (8)	Aufnahmedatum JJJJMMTT
ENTLASS_DATUM	5	DATUM (8)	Entlassungsdatum JJJJMMTT
AUFN_GRUND	6	numerisch (2)	nach Technischer Anlage 2 Schlüssel 1 (1. und 2. Stelle)
AUFN_STATUS	7	numerisch (2)	nach Technischer Anlage 2 Schlüssel 1 (3. und 4. Stelle)
ENTL_GRUND	8	numerisch (2)	nach Technischer Anlage 2 Schlüssel 5 (1. und 2. Stelle)
ENTL_STATUS	9	numerisch (1)	nach Technischer Anlage 2 Schlüssel 5 (3. Stelle)
AUFENTHALT	10	alphanumerisch (1)	Art des Aufenthalts/Behandlung L = vollstationär R = vorstationär N = nachstationär T = teilstationär
DRG	11	alphanumerisch (4)	Vom Krankenhaus gemeldete DRG-Entgelt

Tabelle 13: 301 KH_ICD

Name	ID	Type	Comments
VERSID	1	alphanumerisch	Versichertenpseudonym
ID	2	alphanumerisch	eindeutiger Schlüssel für referentielle Integrität (KH_ICD, KH_OPS)
ICD	3	alphanumerisch (6)	Diagnose ohne Seitigkeit und ohne Sonderkennzeichen
ICD_LOK	4	alphannumerisch (1)	Lokalisation / Diagnose: "L" = links "R“ = rechts "B" = beide
ICD_ART	5	numerisch (1)	Art der ICD-Angabe: 1 = Entlassung/Verlegungsdiagnose, 2 = Hauptdiagnose 3 = Aufnahmediagnose 4 = Einweisungsdiagnose 5= Nebendiagnose
ICD_SEK	6	alphanumerisch (6)	Sekundärdiagnose
ICD_SEK_LOK	7	alphanumerisch (1)	Sekundärdiagnose Lokalisation / Diagnose: "L" = links "R“ = rechts "B" = beide
FACHABT	8	alphanumerisch (4)	Fachabteilung

Tabelle 14: 301 KH_OPS

Name	ID	Type	Comments
VERSID	1	alphanumerisch	Versichertenpseudonym
ID	2	alphanumerisch	eindeutiger Schlüssel für referentielle Integrität (KH_ICD, KH_OPS)

OPS	3	alphanumerisch (8)	OPS ohne Seitenlokalisation
OPS_LOK	4	alphanumerisch (1)	Lokalisation / Diagnose: "L" = links "R“ = rechts "B" = beide
OP_DATUM	5	DATUM (8)	Format: JJJJMMTT

Tabelle 15: 302 Heilm

Name	ID	Type	Comments
VERSID	1	alphanumerisch	Versichertenpseudonym
FALL_ID	2	numerisch	Rezept-ID
BSNR_IK	3	alphanumerisch	Betriebsstättennummer (ggf. alternativ IK des Krankenhauses) aus der Verordnung, pseudonymisiert
LANR	4	alphanumerisch	Lebenslange Arztnummer des Verordners, pseudonymisiert
DATUM_VO	5	DATUM (8)	Verordnungsdatum Format: JJJJMMTT
INDIKATION	6	alphanumerisch	Indikationsschlüssel
VERORDNART	7	numerisch	Kennzeichnung der Art der Heilmittelverordnung (01=Erstverordnung (Regelfall) 02=Folgeverordnung (Regelfall) 10=Verordnung außerhalb des Regelfalles (Folgeverordnung, auch längerfristige Verordnung)
POSNR	8	numerisch (5)	Heilmittelpositionsnummer lt. Heilmittelpositionsnummernkatalog
DATUM_LERB	9	DATUM (8)	Datum der Leistungserbringung (JJJJMMTT)
ANZ_HEI	10	numerisch	Anzahl/Menge der Abrechnungspositionen
FACHGRUPPE	11	alphanumerisch (2)	Fachgruppe des Arztes (bei mehrfachen Fachgruppen je Arzt den Minimum-Wert)

Tabelle 16: 302 Heilm_ICD

Name	ID	Type	Comments
VERSID	1	alphanumerisch	Versichertenpseudonym
FALL_ID	3	numerisch	Rezept-ID
DIAGNOSE	3	alphanumerisch	ICD-Code, Diagnose ohne Seitigkeit und ohne Sonderkennzeichen

Tabelle 17: 302 Hilfsm

Name	ID	Type	Comments
VERSID	1	alphanumerisch	Versichertenpseudonym
BSNR_IK	2	alphanumerisch	Betriebsstättennummer (ggf. alternativ IK des Krankenhauses) aus der Verordnung, pseudonymisiert
LANR	3	alphanumerisch	Lebenslange Arztnummer des Verordners, pseudonymisiert
DATUM_VO	4	DATUM (8)	Verordnungsdatum Format: JJJJMMTT

POSNR	5	numerisch (10)	Hilfsmittelpositionsnummer
ANZ_Hilf	6	numerisch	Anzahl Hilfsmittel
DATUM_LERB	7	DATUM (8)	Datum der Leistungserbringung
FACHGRUPPE	8	alphanumerisch (2)	Fachgruppe des Arztes (bei mehrfachen Fachgruppen je Arzt den Minimum-Wert)

Tabelle 18: XI_PFLEGE

Name	ID	Type	Comments
VERSID	1	alphanumerisch	Versichertenpseudonym
PFL_GR	2	numerisch (1)	Pflegegrad
PFL_GR_VON	3	DATUM (8)	Pflegegrade ab JJJJMMTT
PFL_GR_BIS	4	DATUM (8)	Pflegegrade bis JJJJMMTT
LEI_ART	5	alphanumerisch	Leistungsart Pflege
LEI_VON	6	DATUM (8)	Leistungsbeginn JJJJMMTT
LEI_BIS	7	DATUM (8)	Leistungsende JJJJMMTT

Tabelle 19: IX_REHA_AMB

Name	ID	Type	Comments
VERSID	1	alphanumerisch	Versichertenpseudonym
IK	2	alphanumerisch	BSNR/IK des Leistungserbringers, pseudonymisiert
ID	3	alphanumerisch	Fall-ID
BEGINN	4	DATUM (8)	Datum Beginn JJJJMMTT
ENDE	5	DATUM (8)	Datum Ende JJJJMMTT
ART	6	alphanumerisch	Kategorisierungen wie bei den Krankenkassen genutzt, vorzugsweise im Klartext. Erwartet werden beispielsweise folgende oder vergleichbare Kategorien: ambulante Reha / allgemein ambulante Reha / Atemwegserkrankung ambulante Reha / dermatolog. Erkrankung ambulante Reha / Entwöhnungsbehandlung ambulante Reha / Geriatrie ambulante Reha / integrierte Versorgung ambulante Reha / kardialoglogische Erkrank. ambulante Reha / muskuloskel. Erkrankung ambulante Reha / Neurologische Frühreha ambulante Reha / onkologische Erkrankung ambulante Reha / Psychosomatik
ICD	7	alphanumerisch	Diagnose

Tabelle 20: IX_REHA

Name	ID	Type	Comments
VERSID	1	alphanumerisch	Versichertenpseudonym
IK	2	alphanumerisch	IK der Rehabilitationseinrichtung, pseudonymisiert
ID	3	alphanumerisch	Fall-ID
BEGINN	4	DATUM (8)	Datum Beginn JJJJMMTT
ENDE	5	DATUM (8)	Datum Ende JJJJMMTT
ART	6	alphanumerisch	Kategorisierungen wie bei den Krankenkassen genutzt, vorzugsweise im Klartext. Erwartet werden beispielsweise folgende oder vergleichbare Kategorien: stationäre Anschluss-Reha / AHB stationäre Anschluss-Reha / allgemein stationäre Anschluss-Reha / Atemwegserkrankung stationäre Anschluss-Reha / dermatolog. Erkrankung stationäre Anschluss-Reha / Entwöhnungsbehandlung stationäre Anschluss-Reha / Geriatrie stationäre Anschluss-Reha / integrierte Versorgung stationäre Anschluss-Reha / kardialoglogische Erkrank. stationäre Anschluss-Reha / muskuloskel. Erkrankung stationäre Anschluss-Reha / Neurologische Frühreha stationäre Anschluss-Reha / onkologische Erkrankung stationäre Anschluss-Reha / Psychosomatik stationäre Reha / allgemein stationäre Reha / Atemwegserkrankung stationäre Reha / dermatolog. Erkrankung stationäre Reha / Entwöhnungsbehandlung stationäre Reha / Geriatrie stationäre Reha / integrierte Versorgung stationäre Reha / kardialoglogische Erkrank. stationäre Reha / muskuloskel. Erkrankung stationäre Reha / Neurologische Frühreha stationäre Reha / onkologische Erkrankung stationäre Reha / Psychosomatik
ICD	7	alphanumerisch	Diagnose