Prof. Dr. rer nat. Hendrik Annuth
Mirko Oest BSc
Umut Kurucay BSc
Dr. rer. pol. Heiko Friedel
Mika Santoro BSc
Finn Mertens BSc
Dr. rer. oec. Katja Gehrke

Advanced Analytics in der GKV

KI-gestützte Analyse von Abrechnungsdaten zur Versorgungsoptimierung

Zusammenfassung

Einleitung: Multimorbidität stellt die Versorgung vor Herausforderungen, da komplexe Krankheitsverläufe mit klassischen Verfahren eingeschränkt abgebildet werden. Das Projekt untersucht am Beispiel Diabetes, wie KI-Verfahren in anonymisierten GKV-Routinedaten Patientengruppen sichtbar machen können. Ziel ist eine Beschreibung von Morbiditätsmustern, um Ausreißer und komplexe Versorgungskonstellationen zu erkennen.

Methoden: Verwendet wurden anonymisierte longitudinale GKV-Daten mit ICD-Diagnosen, ATC-Arzneimittelverordnungen, Kostenangaben und soziodemografischen Merkmalen. Die Daten wurden mit neuronalen Netzen in Vektoren überführt. Eingesetzt wurden Embeddings, HDBSCAN, UMAP sowie Pattern-Mining und die Analyse der Vektorvarianz.

Ergebnisse: Die Auswertung zeigt, dass sich multimorbide Patientengruppen differenziert beschreiben lassen. Sichtbar wurden Zusammenhänge zwischen Krankheitsmustern, Ressourcenverbrauch sowie stabilen und dynamischen Verläufen.

Diskussion: Die Ergebnisse verdeutlichen das Potenzial KI-gestützter Analysen und eröffnet neue Ansätze für die Weiterentwicklung der Versorgungsforschung und eine gezieltere Steuerung.

Schlüsselwörter

KI-Verfahren, GKV-Routinedaten, AI-basierte Versorgungsforschung

Crossref/doi

http://doi.org/10.24945/MVF.02.26.1866-0533.2813

Abrechnungsdaten der Gesetzlichen Krankenkassen bilden eine essenzielle Ressource für die Versorgungsforschung, da sie anonymisiert, sektorenübergreifend und standardisiert vorliegen. Traditionelle statistische Analyseverfahren bieten bei steigender Komplexität und Umfang der Daten Vorteile bei der Exploration der Daten. Allein in Deutschland werden jährlich etwa 3,17 Milliarden Diagnosen dokumentiert (Barmer 2024), ergänzt durch weitere umfangreiche Gesundheitsdaten wie Medikations-, Operations- und Abrechnungsinformationen. Künstliche Intelligenz (KI) erweitert – im maschinellen Lernen und mit etablierten statistischen Verfahren – die Analysemöglichkeiten, indem sie nichtlineare Zusammenhänge in großen, vielfältigen Daten erfasst. Praktisch zeigt sich das bei ICD-/ATC-Codes mit sehr vielen Ausprägungen: kompakte, gelernte Zahlenrepräsentationen (Embeddings) verdichten die Kataloge bzw. Codes zu aussagekräftigen Merkmalen und nutzen Ähnlichkeiten sowie Verlaufsinformation. Klassische Alternativen existieren, verlangen jedoch mehr explizite Strukturannahmen und Implementationsaufwand. KI-gestützte Modelle können dadurch eine höhere Prognosegüte erzielen und zusätzliche Perspektiven, z. B. für gezielte Interventionsprogramme zur Vermeidung unnötiger Gesundheitsausgaben eröffnen. Dabei ersetzt KI keineswegs medizinische Expertise, sondern erweitert diese gezielt um datengetriebene Erkenntnisse und Hypothesengenerierung. Ziel dieses Projektes war es, mithilfe moderner KI-Methoden multimorbide Versicherte systematisch zu identifizieren, bei denen die eigentliche Grunderkrankung durch weitere Erkrankungen oder komplexe Krankheitskombinationen überlagert wird. Durch die longitudinale Kombination von Kosten und Daten zu Inanspruchnahmen sollen Ausreißer erkannt und differenzierte Krankheitsprofile erstellt werden.

>> Für diese Studie wurde ein anonymisierter, für Analysezwecke aufbereiteter Auszug aus longitudinalen deutschen Versichertendaten verwendet. Eingeschlossen wurden dabei Versicherte, bei denen im ersten Quartal 2019 (Indexquartal) eine gesicherte Diagnose von Diabetes mellitus Typ 2 (ICD-10: E11*) ambulant oder stationär dokumentiert war. Weiteres Kriterium war, dass sie anschließend entweder mindestens 15 Quartale durchgehend versichert und nachbeobachtbar waren oder im Beobachtungszeitraum (16 Quartale, 01.01.2019 bis 31.12.2022) verstarben.

Für jeden Versicherten wurden quartalsweise soziodemografische Merkmale (Alter, Geschlecht, Sterbedatum), Kosten- und Inanspruchnahmedaten, alle kodierten Diagnosen (ICD-10) sowie Verordnungen (ATC-Codes) extrahiert. Die Kostendaten umfassten ambulante und stationäre Leistungen, Arzneimittel sowie sonstige Leistungen. Die Inanspruchnahme wurde anhand der Anzahl der Arztbesuche, Krankenhausaufenthalte, Hospitalisierungstage und verordneten Arzneimittel pro Quartal erfasst.

Die Abrechnungsdaten enthielten alle pro Quartal bis zu fünfstelligen ICD-10- sowie ATC-Codes, wobei Dubletten entfernt wurden. Die Datenaufbereitung und -reduktion erfolgte mit dem Ziel, die Analysedaten und Modelle kompatibel zum neuen Forschungsdatenzentrum Gesundheit (FDZ) zu gestalten und so künftige Forschungsprojekte mit sensiblen Gesundheitsdaten zu ermöglichen.

Design und Methodik

Das methodische Vorgehen dieser Studie verbindet moderne Verfahren des maschinellen Lernens mit klassischen Ansätzen der Statistik, um komplexe Krankheitsverläufe bei Diabetes mellitus differenziert abzubilden. Die einzelnen Schritte des Analysemodells sind in Abbildung 1 schematisch dargestellt:

Zunächst werden die kodierten Diagnosen (ICD-Codes) und Verordnungen (ATC-Codes) aus den Abrechnungsdaten für jeden Versicherten aggregiert. Anschließend werden diese aggregierten Codes in ausschließlich numerische Werte (Embeddings) transformiert, wodurch die ursprünglich kategorialen Angaben für maschinelle Lernverfahren nutzbar werden. Darüber hinaus erfolgen Vector Embeddings, d. h. eine numerische Darstellung der Daten, die die Bedeutung bzw. Merkmale der Versicherten als Punkte in einem

mehrdimensionalen Raum erfassen. Dadurch können Informationen wie Kosten-, Inanspruchnahme- und soziodemografischen Daten (Alter und Geschlecht) zu einem umfassenden multimodalen Vektor kombiniert werden (Pezeshkpour/Chen 2018).

Schlussendlich werden verschiedene Embedding-Verfahren genutzt, etwa Doc2Vec und Transformer-Modelle, trainiert und anhand von Qualitätsmetriken verglichen, um die leistungsfähigste Variante für die weitere Analyse zu ermitteln.

Abbildung 1 illustriert diese methodische Pipeline – von der Datenaufbereitung, über die Transformation und Evaluierung der Embeddings, das multimodale Feature Engineering bis zur Clusterbildung und Analyse der Krankheitsverläufe.

Zunächst Beginn werden die Versichertenprofile auf Quartalsebene geclustert, anschließend werden die zeitlichen Entwicklungen durch ein zweistufiges Clustering abgebildet. Abschließend erfolgt die Analyse der resultierenden Versichertengruppen, um typische Verläufe, Risikoprofile und auffällige Muster zu identifizieren (Abb. 1).

Übersicht der methodischen Schritte von der Datenaufbereitung bis zur Clusteranalyse

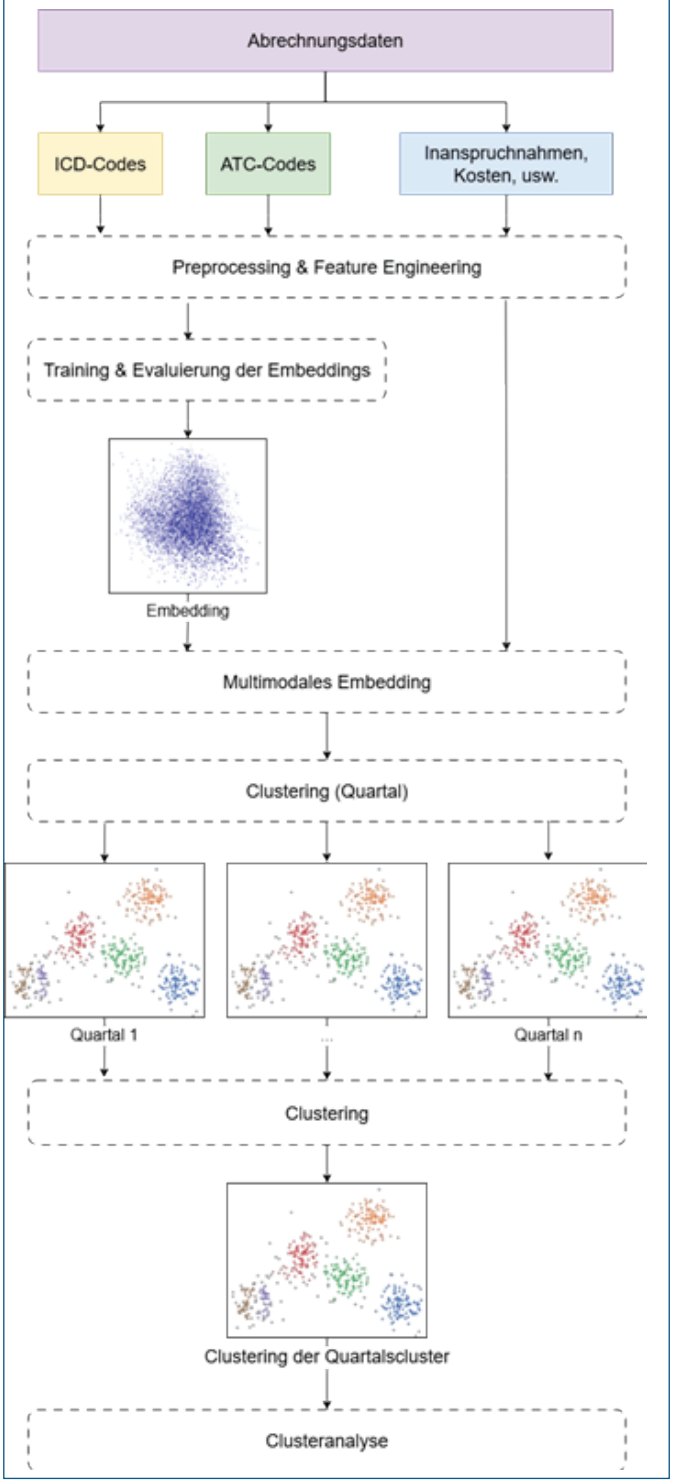


Abb. 1: Übersicht der methodischen Schritte von der Datenaufbereitung bis zur Clusteranalyse. Eigene Darstellung.

Datenaufbereitung und Feature Engineering

Die für diese Studie verwendeten Abrechnungsdaten bestehen überwiegend aus nicht-numerischen, variierenden ICD- und ATC-Codes. Damit diese Daten für maschinelle Lernverfahren genutzt werden können, ist eine numerische Repräsentation fester Größe erforderlich, die die klinische Bedeutung und Zusammenhänge zwischen den Codes erhält (Pezeshkpour/Chen 2018). Dafür werden alle über 16 Quartale dokumentierten ICD- und ATC-Codes je Versicherten aggregiert. Das resultierende Profil umfasst sämtliche Diagnosen und Medikationen der letzten vier Jahre und bildet durch wiederholtes Auftreten chronische oder persistierende Erkrankungen ab (Steiger/Kroll 2023).

Durch Verwendung der o. g. Embeddings liegen Versicherte mit ähnlichen Krankheitsbildern auch im Zahlenraum eng beieinander. Damit lassen sich sowohl klinische Ähnlichkeiten als auch Wechselwirkungen zwischen Diagnosen und Medikamenten abbilden (Abb. 2).

Für die Embedding-Erzeugung wird das bewährte Doc2Vec-Modell genutzt, das bereits erfolgreich auf Versichertendaten angewendet wurde (Le/Mikolov 2014; Steiger/Kroll 2023). Eingesetzt wird die DBOW-Variante, bei der das gemeinsame Auftreten von Codes – nicht deren Reihenfolge – im Mittelpunkt steht.

Für die Feinabstimmung der Modellparameter kommt eine Bayesian Optimization zum Einsatz (Frazier 2018). Zur Bewertung der Embedding-Qualität dient der Rank Match Score (RMS), der misst, wie gut das Modell diagnostisch ähnliche Patienten im numerischen Raum korrekt gruppiert (Taskesen 2025; Kendall 1938). Die Berechnung erfolgt durch Vergleich der räumlichen Distanz im Embedding-Space mit der Ähnlichkeit, etwa mittels Jaccard-Koeffizient. Zusätzlich wird ein vortrainiertes Pat2Vec-Modell (Steiger/Kroll 2023) für den aktuellen Datensatz angepasst und gegen das eigene Modell validiert.

Das Ergebnis ist ein multimodaler Vektor pro Versicherten, der medizinische, demografische und versorgungsbezogene Merkmale integriert und als Grundlage für alle weiteren Analysen dient.

Evaluierung der Embeddings

Die Qualität der erzeugten ICD- und ATC-Code-Embeddings wurde durch verschiedene komplementäre Methoden überprüft, um sowohl die interne Struktur als auch den praktischen Nutzen der Embeddings zu beurteilen. Die Evaluierung basierte auf einer großen Ko-

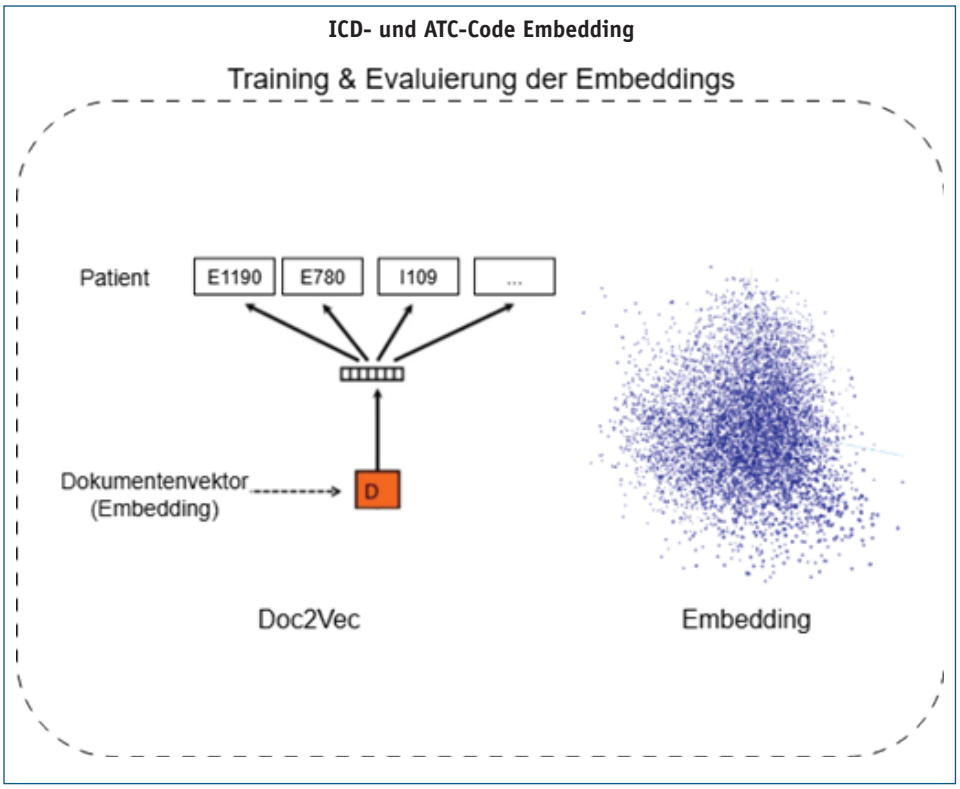


Abb. 2: ICD- und ATC-Code Embedding. Eigene Darstellung.

horte von 341.887 Patienten, für die sämtliche ICD- und ATC-Codes über 16 Quartale analysiert wurden. Untersucht wurden sowohl Embeddings, die ausschließlich ICD-Codes als auch solche, die ICD- und ATC-Codes gemeinsam abbilden.

Im Rahmen der intrinsischen Evaluierung wurde geprüft, wie gut die Embeddings die inhaltliche Nähe zwischen medizinischen Codes und Versichertenprofilen widerspiegeln. Hierbei kam der Neighborhood Overlap Score zum Einsatz, der misst, inwieweit Nachbarschaften im Embedding-Raum den klinischen Ähnlichkeiten entsprechen. Der Wert reicht von 0 (keine Übereinstimmung) bis 1 (perfekte Übereinstimmung); höhere Werte sprechen für eine bessere semantische Konsistenz (Chen/Buja 2009; Lin/Smith 2019; Flexer 2015).

Für die extrinsische Evaluierung wird geprüft, ob die Repräsentationen der Daten tatsächlich in praktischen Anwendungsszenarien helfen. Als prototypische Aufgabe wurde hier die Vorhersage des Geschlechts eines Patienten (Klassifikation) und die aggregierten Gesundheitskosten (Regression) gewählt. Diese wurde dann jeweils mit einem klassischen Random Forest Ansatz vorhergesagt. Der Zweck dieser Betrachtung ist der Erhalt der wesentlichen Informationen innerhalb der Daten nach dem Embedding. Die Modellgüte wurde für das Geschlecht durch Accuracy und AUROC bewertet (Steiger/Kroll 2023) für das Regressionsmodell Mean Absolute Error (MAE) und R^2 als Bewertungsgrößen genutzt wurden (Steiger/Kroll 2023). Ein niedriger MAE und hoher R^2-Wert sprechen für eine gute Modellqualität und hochwertige Embeddings.

Durch die Kombination dieser intrinsischen und extrinsischen Bewertungsverfahren konnte sowohl die interne Struktur als auch die praktische Anwendbarkeit der Embeddings umfassend geprüft werden.

Zweistufiges Clustering der multimodalen Embeddings

Das Clustering der multimodalen Versicherten-Embeddings erfolgte in einem zweistufigen Verfahren, um sowohl Einzelzeitpunkte als auch die zeitliche Entwicklung der Patientenverläufe zu berücksichtigen. In der ersten Stufe wurden die individuellen Embeddings der Versicherten für jedes Quartal separat mittels HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) gruppiert. Es handelt sich um ein dichtebasiertes Clustering-Verfahren, das Gruppen (Cluster) von Datenpunkten identifiziert, ohne dass die Anzahl der Cluster vorher festgelegt werden muss. Zur besseren Strukturierung der hochdimensionalen Daten wurde zuvor UMAP (Uniform Manifold Approximation and Projection) als nichtlinearer Algorithmus zur Dimensionsreduktion eingesetzt (Steiger/Kroll 2023; Malzer/Baum 2021). In der zweiten Stufe wurden die über alle Quartale hinweg ermittelten Clusterzuordnungen eines Patienten zu einem Trajektorien-Vektor (zeitlichen Verlauf eines Objekts oder Systems durch den Raum) zusammengefasst, bei dem jede Dimension ein Quartal repräsentiert. Diese Vektoren wurden erneut mit HDBSCAN, diesmal ohne UMAP, unter Verwendung der Hamming-Distanz geclustert, da sie Unterschiede in kategorialen Werten optimal abbildet (Allaoui/Kherfi 2020).

Dieses zweistufige Clustering ermöglicht es, nicht nur Momentaufnahmen zu analysieren, sondern auch die Dynamik und Entwicklung von Krankheits- und Versichertenpfaden über den Beobachtungszeitraum hinweg differenziert abzubilden und Versichertengruppen mit ähnlichen Verlaufsmustern gezielt zu identifizieren.

Clusteranalyse

Um die charakteristischen Merkmale für die Zuordnung eines Versicherten zu einem bestimmten Cluster zu identifizieren, wurde eine umfassende Clusteranalyse durchgeführt. Ziel war es, die Versichertengruppen klar voneinander abzugrenzen und für jedes Cluster ein eigenständiges Profil hinsichtlich Krankheitslast, Versorgung und Ressourcenverbrauch zu erarbeiten.

Das Embedding der Versichertendaten war hierbei zentral, da es die Vielzahl nicht-numerischer Merkmale in eine vergleichbare numerische Form überführt und gleichzeitig die semantischen Zusammenhänge zwischen Diagnosen und Medikationen bewahrt. Erst dadurch war eine sinnvolle Gruppierung auf Basis komplexer medizinischer Muster möglich.

Für quantitative Merkmale wie Kosten und Inanspruchnahme wurden klassische deskriptive Statistiken wie Mittelwert und Standardabweichung berechnet, um die Verteilung innerhalb der Cluster zu analysieren und typische Versorgungsprofile zu erkennen. Für

die zahlreichen diagnostischen Merkmale (ICD- und ATC-Codes), die oft in komplexen Kombinationen auftreten, kam Pattern-Mining zum Einsatz (Heaton 2017). Dadurch ließen sich sowohl häufige Einzelcodes als auch charakteristische Kombinationen und Muster innerhalb der Cluster identifizieren. So konnte beispielsweise ein Cluster durch das gehäufte gemeinsame Auftreten von Diabetes (ICD E11), Hypertonie (ICD I10) und bestimmten Lipidsenker-Verordnungen (ATC C10AA) charakterisiert werden. Der Vergleich solcher Muster zwischen den Clustern ermöglichte die Ableitung cluster-spezifischer Krankheits- und/oder Medikationsprofile.

Zur weitergehenden Charakterisierung wurde die sogenannte Vektorvarianz berechnet. Sie misst die durchschnittliche Streuung der Merkmalsvektoren eines Patienten über alle betrachteten Quartale. Eine geringe Varianz steht für stabile, gleichbleibende Krankheitsverläufe, während eine hohe Varianz auf dynamische oder wechselhafte Verläufe und Therapieansätze hindeutet.

Übersicht der Leistungsmetriken aller getesteten Embeddings

Metrik/Modell	Overlap Score	Klassifikation Accuracy	Klassifikation AUC	Regression MAE	R^2
Random/ Referenzmodell	0,53	0,54	0,50	22.626	0,00
Pat2Vec	0,61	0,77	0,84	18.831	0,20
Doc2Vec Dim 60	0,62	0,82	0,89	19.502	0,27
Custom BERT	0,69	0,82	0,90	18.295	0,26

Tab. 1: Übersicht der Leistungsmetriken aller getesteten Embeddings. Hinweis: MAE: niedriger ist besser; alle übrigen Metriken: höher ist besser. Eigene Darstellung.

Ergebnisse

Die Analysen ermöglichten eine differenzierte Charakterisierung von Diabetespatienten und legten komplexe Muster ihrer Krankheitsverläufe, Komorbiditäten und Ressourceninanspruchnahme offen. Die Ergebnisse betreffen die Evaluation der Embedding-Modelle, die Beschreibung der identifizierten Patientengruppen durch Clustering sowie die Analyse charakteristischer medizinischer Muster.

Modellentwicklung und Evaluierung der Patienten-Embeddings

Zur Umwandlung der komplexen, nicht-numerischen Patientendaten (ICD- und ATC-Codes) wurden verschiedene Embedding-Ansätze entwickelt und verglichen, darunter das Referenzmodell Pat2Vec, ein zufallsbasiertes Null-Modell, mehrere Doc2Vec-Varianten (inklusive ICD-only- und hochdimensionalen Modellen) sowie ein eigens trainiertes Transformer-Modell (Custom BERT). Die Leistungsbewertung erfolgte anhand von Metriken wie Overlap Score, Klassifikationsgenauigkeit (Accuracy), AUROC, mittlerer absoluter Fehler (MAE) und Bestimmtheitsmaß (R^2).

Das Referenzmodell zeigte die schwächsten Ergebnisse (Overlap Score 0,53; Accuracy 54%; AUROC 0,50; MAE 22.626 Euro; R^2 0,00). Das Pat2Vec-Modell war mit Overlap Score 0,61, Accuracy 77 %, AUROC 0,84, MAE 18.831 Euro und R^2 0,20 bereits deutlich besser. Die größte Doc2Vec-Variante (60 Dimensionen) erzielte eine Accuracy von 82%, AUROC 0,89, Overlap Score 0,62, MAE 19.502 Euro und mit R^2 0,27 den höchsten Wert aller Modelle. Das Custom BERT-Modell übertraf mit Overlap Score 0,69, Accuracy 82%, AUROC 0,90, MAE 18.295 Euro und R^2 0,26 alle anderen Modelle.

Insgesamt zeigte das Transformer-Modell (Islam/Saidul 2023) die beste Performance über alle Qualitätsmetriken hinweg, insbesondere bei der Klassifikation des Patientengeschlechts und der Kostenprognose. Auch der Overlap Score, der die Fähigkeit zur Gruppierung ähnlicher Patienten misst, war hier am höchsten. Das Pat2Vec-Modell diente als Referenz für die Grundvalidierung des Ansatzes, aber die Experimente mit modernen Embedding-Methoden verdeutlichen das Potenzial der Transformer-Technologie für noch präzisere Analysen und verbesserte Clustering-Ergebnisse (Tab. 1).

Patientencharakterisierung durch Clustering-Analyse

Die auf den multimodalen Embeddings basierende, zweistufige HDBSCAN-Clusteranalyse identifizierte innerhalb der Kohorte von über 340.000 Diabetikern insgesamt 14 Cluster. Die sieben aussagekräftigsten Cluster lassen sich in fünf Hauptgruppen einteilen: Cluster mit stabilen Diabetesverlauf, Hochrisikocluster, Cluster mit hoher Multimorbidität und Mortalität, Cluster hoher Arzneimittelkosten, sowie in ein großes Cluster mit hoher Variabilität.

Die beiden größten Cluster (7 und 13) umfassen etwa 131.000 Patienten mit stabilem Diabetesverlauf, niedrigem Durchschnittsalter (68 bzw. 65 Jahre), eher geringer Vektorvarianz, sehr niedrigen Kosten (600 Euro pro Jahr) und 100 % Überlebensrate. Die Hochrisikocluster 9 und 10 beinhalten ca. 14.000 Patienten mit deutlich höherem Alter (79 bzw. 81 Jahre), hoher Mortalität (nur 18 % bzw. 4 % Überlebende), vergleichsweise langen Krankenhausaufenthalten (10,5 bzw. 12,1 Tage) und höchsten Kosten (8.600 bzw. 10.100 Euro). Das Cluster 4 (17.000 Patienten, Alter 81) steht für hohe Multimorbidität und mittlere Mortalität (54 % Überlebende). Das Cluster 1 (ca. 3.900 Patienten) ist durch besonders hohe Arzneimittelausgaben auffällig. Der Subpopulation mit hoher Variabilität (–1) mit 133.000 Patienten weist eine relativ hohe Vektorvarianz (13,5), ein mittleres Alter (72 Jahre), heterogene Krankheitsverläufe und eine Überlebensrate von 87 % auf. Insgesamt zeigt sich eine große Heterogenität in der Diabetikerpopulation, wobei eher stabile, hochkomplexe und instabile Verläufe ebenso wie besondere „Ausreißergruppen" datenbasiert identifiziert werden können.

Eine niedrige Vektorvarianz weist darauf hin, dass die Werte innerhalb des Vektorraums nur geringe Abweichungen vom Mittelwert zeigen. Dies bedeutet, dass die Vektorkomponenten eng um den Mittelwert gruppiert sind und nur eine geringe Dispersion also Streuung aufweisen.

Medizinische Muster und charakteristische Krankheitskombinationen:

Die detaillierte Clusteranalyse ermöglichte die Identifikation häufiger und spezifischer ICD- und ATC-Code-Kombinationen. Besonders bei typischen Diabetikern (Cluster 7 und 13) trat die bekannte Trias aus Diabetes mellitus Typ 2 (E11.9), essenzieller Hypertonie (I10.9) und reiner Hypercholesterinämie (E78.0) auf. Weitere Kombinationen wie Diabetes mit Hyperlipidämie (E78.5)

Beschreibung der aussagekräftigsten Cluster (Auszug Sieben von 14)							
Gruppe	**stabiler Diabetesverlauf**		**Hochrisiko-Cluster**		**Hohe Multimorbidität/ Mortalität**	**Hohe Arzneimittelkosten**	**Subpopulation mit hoher Variabilität**
Cluster	7	13	9	10	4	1	-1
Patientenanzahl (Tsd.)	60	71	6	8	17	4	133
Altersdurchschnitt	68	65	79	81	81	69	72
Anteil lebendig	1	1	0,18	0,04	0,54	1	0,87
Vektorvarianz	6,6	6,2	21	0,8	11,4	11,3	13,5
KH-Tage Ø	0,1	0,1	10,5	12,1	1,9	1,7	2,7
Gesamtkosten Ø (Tsd.)	**0,6**	**0,6**	**8,6**	**10,1**	**2,6**	**2,7**	**3,4**

Tab. 2: Beschreibung der aussagekräftigsten Cluster (Auszug Sieben von 14). Eigene Darstellung.

und Hyperurikämie (E79.0) wiesen auf das metabolische Syndrom hin (Alberti et al. 2009). Daneben waren Komorbiditäten wie Gonarthrose (M17.9), atherosklerotische Herzkrankheit (I25.1) (Zhang et al. 2016), Presbyopie (H52.2) und Astigmatismus (H52.4) typisch (Klein/Klein 1997). In den Hochrisikoclustern wurden seltene, aber klinisch besonders relevante Konstellationen gefunden, etwa Diabetes in Kombination mit Hypokaliämie (E87.6) und akuter respiratorischer Insuffizienz (J96.0), mit Linksherzinsuffizienz (I50.1) und akutem Nierenversagen (N17.9) oder akuter Blutungsanämie (D62) mit Diabetes und Hypertonie.

Diese Ergebnisse zeigen, dass die KI-gestützte Analyse von Abrechnungsdaten sowohl typische als auch seltene und klinisch bedeutsame Muster datenbasiert erfassen kann – eine wesentliche Grundlage für gezielte Versorgungs- und Interventionsstrategien.

Diskussion und Anwendungsbedeutung

Die Ergebnisse dieser Studie zeigen, dass moderne KI-basierte Methoden erstmals ermöglichen, große Routinedatenbestände wie GKV-Abrechnungsdaten datengetrieben und ohne die Notwendigkeit der Integration etwa klinischer Daten differenziert auszuwerten. Durch die Kombination von Kosten-, Inanspruchnahmedaten einschließlich Diagnosen und verschriebenen Arzneimitteln in multimodalen Feature-Embeddings sowie einen zweistufigen Clustering-Ansatz wurde ein innovatives Verfahren zur Identifikation von Versichertenmuster entwickelt. Besonders dort, wo klinische Zusatzinformationen fehlen, eröffnet dieses Vorgehen neue Möglichkeiten für Auswertung und Steuerung der Versorgung. Die Methode eignet sich sowohl zur Identifikation von Besonderheiten in den Versichertenprofilen und komplexen Patientenverläufen als auch zur Charakterisierung typischer Krankheitsverläufe über die Zeit.

Das Verfahren kann epidemiologische Studien verbessern, indem Verzerrungen durch überlagernde Erkrankungen berücksichtigt werden. Darüber hinaus könnte man durch diese Art der Segmentierung gezielter Zuz- und Abschläge im Morbi-RSA berücksichtigen. Oder Krankenkassen könnten den Algorithmus intern verwenden, um passgenauere Angebote für die Versicherten zu formulieren. Letzteres setzt allerdings voraus, dass GKV-Abrechnungsdaten schneller als bislang für Auswertungen zur Verfügung stehen. Gleichzeitig zeigt sich, dass die datengetriebene Identifikation von unterschiedlichen Versichertenprofilen einer stärkeren Integration medizinischen Expertenwissens profitieren würde. Die Entwicklung von Referenzdatensätzen, in denen „Ausreißer" und typische Patienten fachlich klassifiziert sind, könnte die Aussagekraft künftiger Analysen und Klassifikationsmodelle weiter erhöhen.

Fazit

Das Projekt belegt, wie KI-gestützte Methoden genutzt werden können, um in Routinedaten komplexe multimorbide Patientengruppen zu identifizieren, deren Grunderkrankung durch andere Diagnosen oder Krankheitskombinationen überdeckt ist. Grundlage bildeten quartalsbezogene Diagnosen (ICD-Codes) und Arzneimittelverordnungen (ATC-Codes), die mittels neuronaler Netze in numerische Embeddings transformiert wurden. Diese Vektoren

Literatur

1. BARMER Ersatzkasse. Arztreport 2024. Wuppertal: BARMER; 2024. S. 74. Verfügbar unter: https://www.barmer.de/resource/blob/1255894/d58c1bd7c1a99c78a54d410c490920c8/dl-arztreport-2024-data.pdf#page=74. Zugegriffen: 19. Jun. 2025.
2. Kassenärztliche Bundesvereinigung. Einheitlicher Bewertungsmaßstab (EBM), Stand: 2. Quartal 2025 [Internet]. 2025. Verfügbar unter: https://www.kbv.de/html/ebm.php. Zugegriffen: 19. Jun. 2025.
3. Alberti et al. (2009): Harmonizing the Metabolic Syndrome. Circulation.
4. Klein R, Klein BE. "Diabetic eye disease." Lancet. 1997;350 Suppl 1:SII20-3.
5. Zhang et al. (2016): Association between diabetes mellitus and osteoarthritis: systematic review and meta-analysis. Osteoarthritis Cartilage.
6. Pezeshkpour et al. (2018): Embedding Multimodal Relational Data for Knowledge Base Completion. University of California, Irvine.
7. Le & Mikolov (2014): Distributed Representations of Sentences and Documents. Google Inc.
8. Frazier (2018): A Tutorial on Bayesian Optimization.
9. Taskesen (2025): K-Nearest-Rank Scores: A Tool for Local Structure Analysis in Embeddings. Data Science Collective, Medium.
10. Kendall (1938): A New Measure of Rank Correlation. Biometrika, Volume 30, Issue 1-2, Pages 81–93.
11. Malzer & Baum (2021): A Hybrid Approach To Hierarchical Density-based Cluster Selection.
12. Steiger & Kroll (2023): Patient Embeddings From Diagnosis Codes for Health Care Prediction Tasks: Pat2Vec Machine Learning Framework.
13. Allaoui et al. (2020): Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study.
14. Heaton (2017): Comparing Dataset Characteristics that Favor the Apriori, Eclat or FP-Growth Frequent Itemset Mining Algorithms. College of Engineering and Computing.
15. Chen & Buja (2009): Local Multidimensional Scaling for Nonlinear Dimension Reduction, Graph Drawing, and Proximity Analysis. Journal of the American Statistical Association.
16. Lin & Smith (2019): Situating Sentence Embedders with Nearest Neighbor Overlap (N2O).
17. Flexer (2015): Improving Visualization of High-Dimensional Music Similarity Spaces.
18. Islam et al. (2023): A Comprehensive Survey on Applications of Transformers for Deep Learning Tasks.

ermöglichen die datenbasierte Gruppierung von Patienten und die Analyse von Beziehungen zwischen medizinischen Merkmalen und Ressourcenverbrauch.

Das zweistufige HDBSCAN-Clustering – ergänzt durch UMAP – erfasst nicht nur Momentaufnahmen, sondern auch zeitliche Verläufe. Die Integration von Pattern-Mining und die Bewertung der Vektorvarianz liefern zusätzliche Erkenntnisse über die Stabilität oder Dynamik von Krankheitsbildern innerhalb der Cluster. Die Ergebnisse zeigen, dass KI-gestützte Analysen anonymisierter Abrechnungsdaten effektiv zur Identifikation von Ausreißern und komplexen Versorgungskonstellationen beitragen können. Die Methodik bietet einen Rahmen zur datenbasierten Weiterentwicklung der Versorgungsforschung und -steuerung und bildet die Grundlage für zukünftige Arbeiten, in denen durch gezielte Integration von Expertenwissen und weiteren Kontextdaten die praktische Relevanz noch weiter gesteigert werden kann. <<

Advanced analytics in the Statutory Health Insurance system: AI-supported analysis of billing data for healthcare optimization

Introduction: Multimorbidity poses challenges for healthcare delivery, as traditional methods have limited ability to capture complex disease trajectories. Using diabetes as a case study, this project investigates how AI methods can identify patient groups within anonymized routine statutory health insurance data. The goal is to describe patterns of morbidity in order to identify outliers and complex care constellations.

Methods: We used anonymized longitudinal statutory health insurance data including ICD diagnoses, ATC drug prescriptions, cost data, and sociodemographic characteristics. The data were converted into vectors using neural networks. We employed embeddings, HDBSCAN, UMAP, pattern mining, and vector variance analysis.

Results: The analysis shows that multimorbid patient groups can be described in a differentiated manner. Correlations became apparent between disease patterns, resource utilization, and stable and dynamic disease courses.

Discussion: The results highlight the potential of AI-supported analyses and open up new approaches for the further development of health services research and more targeted management.

Keywords

AI-methods, SHI-billing data, AI-based healthcare research

Autorenerklärung

Die Autor:innen erklären, dass keine Interessenskonflikte vorliegen.

Zitationshinweis

Annuth et al.: „Advanced Analytics in der Gesetzlichen Krankenversicherung: KI-gestützte Analyse von Abrechnungsdaten zur Versorgungsoptimierung", in: „Online First zu Monitor Versorgungsforschung" (02/26), S. 57–63. http://doi.org/10.24945/MVF.02.26.1866-0533.2814

Prof. Dr. rer nat. Hendrik Annuth ORCID: 0009-0007-1797-8063
übernahm 2020 den Lehrstuhl für Data Science & Artificial Intelligence an der Fachhochschule Wedel und lehrt die Veranstaltungen Data Literacy, Exploratory Data Analysis, Machine Learning und Deep Learning. Er ist Studiengangsleiter sowohl für den Bachelor- als auch den Masterstudiengang und betreut im Rahmen seiner Lehrtätigkeit zahlreiche praxisrelevante Studienprojekte. Kontakt: Hendrik.Annuth@fh-wedel.de



Mirko Oest BSc ORCID: 0009-0000-3532-2737
hat einen Bachelorabschluss im Studiengang Wirtschaftsingenieurwesen an der FH Wedel. Er setzt sein Studium im Masterprogramm „Data Science & Artificial Intelligence" an derselben Hochschule fort. Seine akademische Laufbahn verbindet ingenieurwissenschaftliches Know-how mit modernen datenwissenschaftlichen Methoden. Kontakt: MirkoOest@googlemail.com



Umut Kurucay BSc ORCID: 0009-0004-3543-0227
ist dualer Masterstudent bei Airbus mit Fokus auf datenbasierte Anwendungen in der Luftfahrt. Davor war er als .NET-Entwickler bei best4bps tätig und absolvierte ein duales Bachelorstudium bei der Deutschen Telekom mit dem Schwerpunkt auf Prozessautomatisierung. Er studierte Wirtschaftsinformatik an der FHDW Bielefeld und befindet sich aktuell im Masterstudium „Künstliche Intelligenz & Data Science" an der FH Wedel. Kontakt: hello@umutkurucay.com



Dr. rer. pol. Heiko Friedel
ist im Bereich der Versorgungsforschung und Gesundheitsökonomie tätig. Er ist u. a. verantwortlich für die Durchführung von wissenschaftlichen Studien und Projekten sowie für Gutachten zur Beurteilung der Wirtschaftlichkeit verschiedener Leistungen der Gesetzlichen Krankenkassen.
Kontakt: Heiko.Friedel@teamgesundheit.de



Mika Santoro BSc ORCID: 0009-0003-5631-5423
ist dualer Masterstudent bei Airbus. Davor absolvierte er ein duales Bachelorstudium bei Hewlett Packard Enterprise (HPE) mit der Vertiefung „Computational Data Science". Er studierte Informatik an der DHBW Stuttgart und befindet sich aktuell im Masterstudiengang „Data Science & Artificial Intelligence" an der FH Wedel.
Kontakt: Mika.Santoro@gmail.com



Finn Mertens BSc ORCID: 0009-0003-8124-7856
ist Masterstudent im Bereich Informatik an der Fachhochschule Wedel. Außerdem ist er werkstudentisch als Software Developer bei adesso SE tätig. Zuvor arbeitete er als werkstudentisch der Software Quality Assurance bei Diebold Nixdorf und als IT Security Consultant bei KPMG.
Kontakt: inf103461@stud.fh-wedel.de



Dr. rer. oec. Katja Gehrke ORCID: 0000-0001-5516-457X
ist Geschäftsführende Wissenschaftlerin des Instituts für deskriptive Gesundheitsdatenanalyse (DGDA). Davor arbeitete sie in leitender Funktion beim Bundesverband der Pharmazeutischen Industrie, u.a. mit dem Fokus Gesundheitsökonomie und Versorgungsforschung sowie Strategie und Marktzugang. Sie studierte Volkswirtschaftslehre und war am Lehrstuhl für Finanzwissenschaft und Gesundheitsökonomie der TU Berlin und an der L'Université libre de Bruxelles (ULB) tätig. Kontakt: KGehrke@dgda-institut.de

