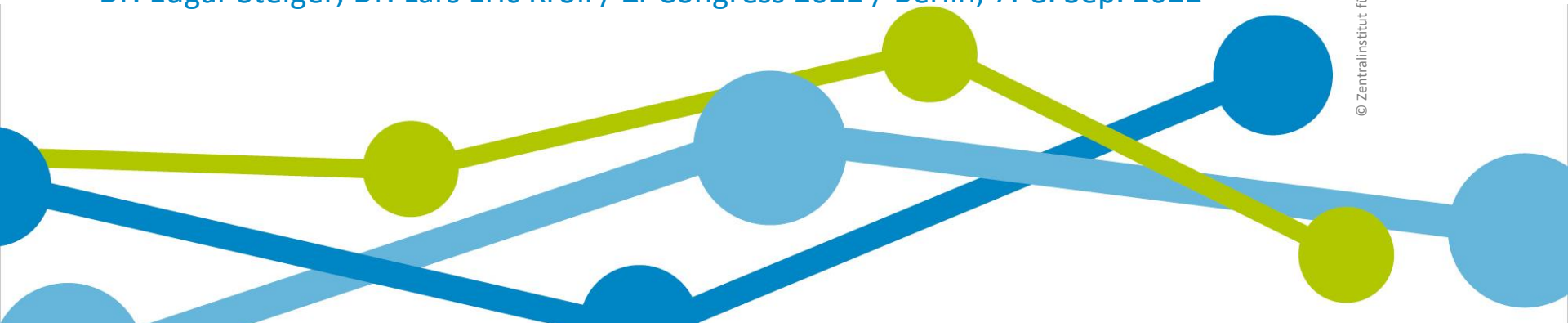
ZENTRALINSTITUT FÜR DIE
KASSENÄRZTLICHE VERSORGUNG
IN DEUTSCHLAND

Potenziale für die Optimierung von Versorgungsprozessen mithilfe von Machine Learning auf Abrechnungsdaten **Pat2Vec**

Dr. Edgar Steiger, Dr. Lars Eric Kroll / Zi-Congress 2022 / Berlin, 7.-8. Sep. 2022



Stat. Modelle und Machine Learning auf Abrechnungsdaten
bisher: Binärkodierung ausgewählter Diagnosen

Pat.	JQ	Diagnose-code
Pat. 1	20141	E11.9
Pat. 1	20153	F32.0
Pat. 2	20144	J01.8
Pat. 3	20134	J01.9
Pat. 3	20152	I10.9
…	…	…

Pat.	E11.9	F32.0	J01	I10.9	…
Pat. 1	1	1	0	0	…
Pat. 2	0	0	1	0	…
Pat. 3	0	0	1	1	…
…	…	…	…	…	…

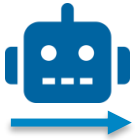
Prädiktion

Pat.	Y
Pat. 1	101,3
Pat. 2	47,5
Pat. 3	16,7
…	…

Charakterisierung
ohne Outcome,
z.B. Clustering

Ansatz „Pat2Vec“: Vektorrepräsentation aller Diagnosen

Pat.	Alle Codes
Pat. 1	„E11.9 F32.0 E11.20 …“
Pat. 2	„J01.8 O22.4 O80 …“
Pat. 3	„J01.9 I10.9 E66.99 F32.0 …“
…	…



Pat.	Dim1	Dim2	Dim3	…
Pat. 1	0,5	-0,1	0,3	…
Pat. 2	-0,3	0,8	-1,1	…
Pat. 3	1,4	0,1	0,4	…
…	…	…	…	…

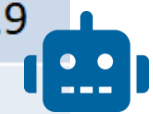
Prädiktion

Pat.	Y
Pat. 1	101,3
Pat. 2	47,5
Pat. 3	16,7
…	…

Charakterisierung ohne Outcome, z.B. Clustering

Ziel: vollständige Diagnoseinformationen nutzbar machen

Pat.	JQ	Diagnose-code
Pat. 1	20141	E11.9
Pat. 1	20153	F32.0
Pat. 2	20144	J01.8
Pat. 3	20134	J01.9
Pat. 3	20152	I10.9
…	…	…



- Verbesserte Prädiktion von **allgemeinen** versorgungsrelevanten Outcomes
- Differenziertere Charakterisierung von Patient*innen bzgl. ihres **Versorgungsbedarfs**
- Potenzial von fortgeschrittenen **Machine-Learning**-Ansätzen erkunden

Herausforderungen Vektorisierung

- Vektorisierung v.a. zur Beschreibung des **allgemeinen Krankheitsprofils/Versorgungsbedarfs** (und nicht eines speziellen Outcomes wie z.B. „Knochenbruch“)
- Vektorisierungsalgorithmus hat mehrere festzulegende **Hyperparameter**
- **Güte der Vektorisierung** lässt sich nicht direkt messen

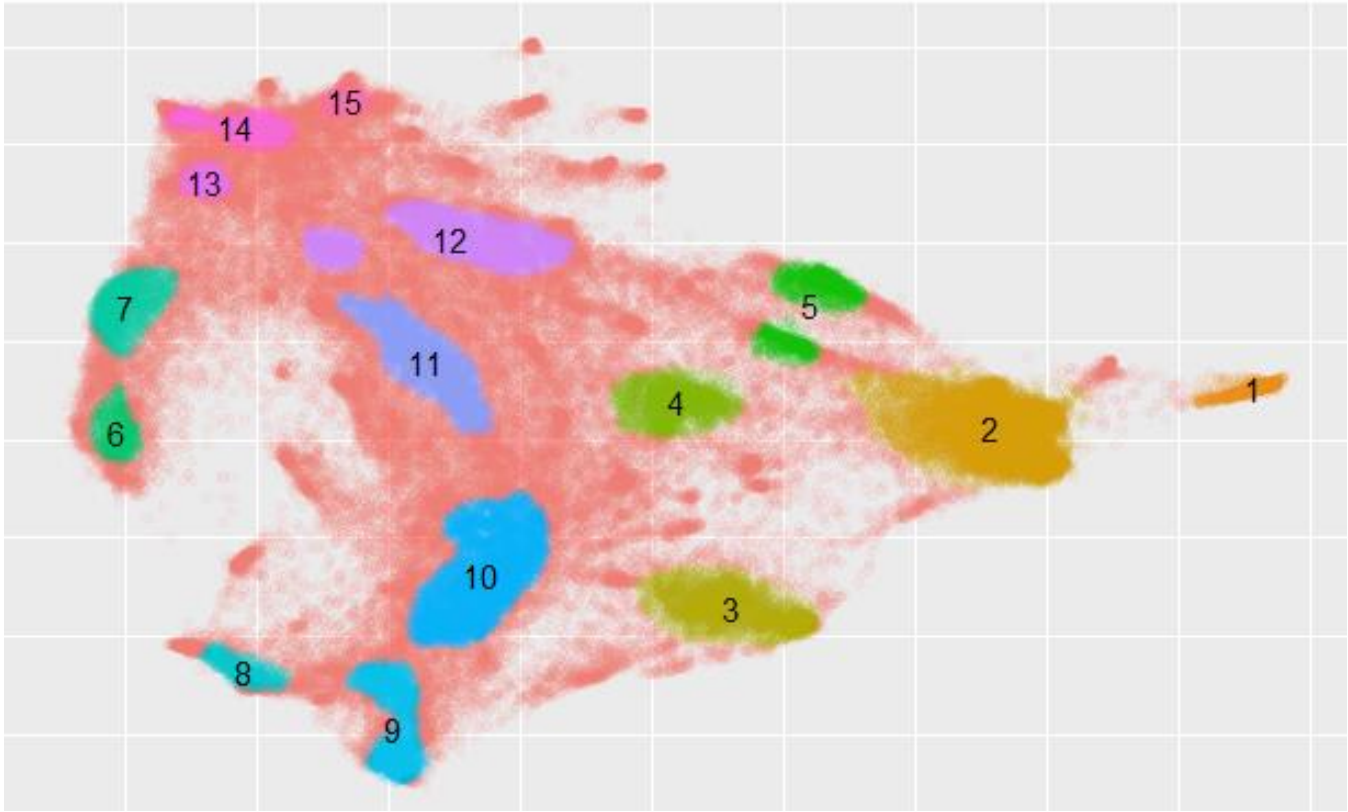
Lösung:

Machine-Learning-Pipeline zum Finden der besten Hyperparameter/des besten allgemeinen Vektorisierungsmodells mit Hilfe eines konstruierten aggregierten Gütemaßes

Methodik

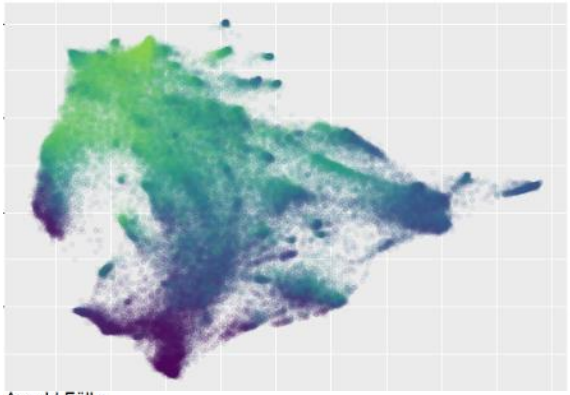
- **Doc2Vec** auf stringbasierten Diagnosedaten von 11,2 Mio. Patient*innen
- Aggregierter Gesamtscore
  - allgemeine Prädiktionsaufgaben (Alter, Geschlecht, Fallzahl, Notfallstatus)
  - verschiedene Maßzahlen (AUROC, R2, …)
  - Verschiedene Algorithmen (lin./log. Regression sowie Boosted Trees)
- Bayessche Optimierung der Hyperparameter von Doc2Vec
- Evaluation des finalen Modells mit neuer Prädiktionsaufgabe (Verordnungskosten) sowie grafische Darstellung mit UMAP
- Ausführliche Darstellung im **Preprint** under Review: https://preprints.jmir.org/preprint/40755

Visualisierung: Pat.-Cluster in der Vektorisierung

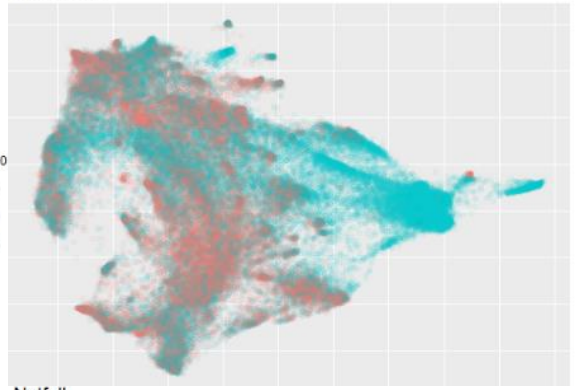


Visualisierung: Eigenschaften der Patient*innen

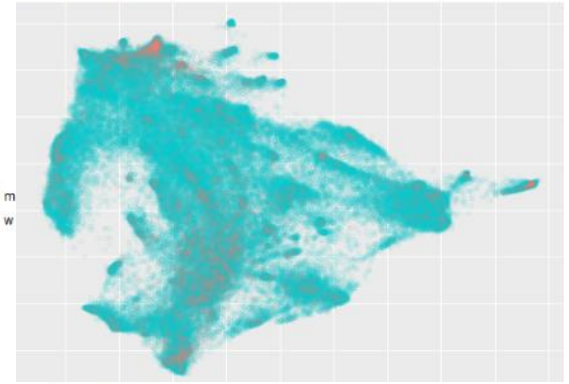
Alter



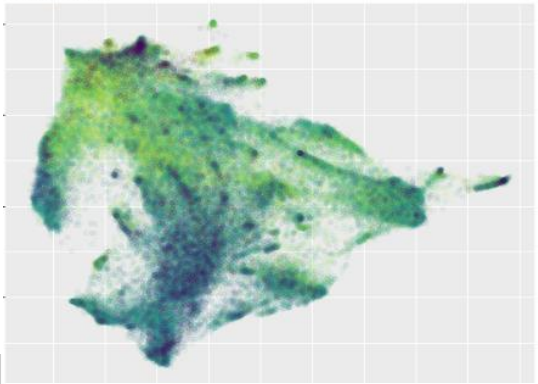
Geschlecht



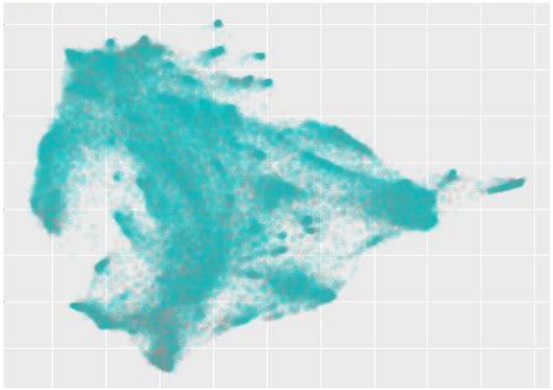
letztes Jahr in Daten



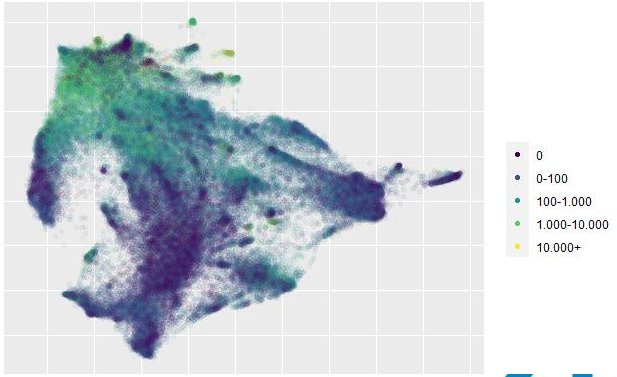
Anzahl Fälle



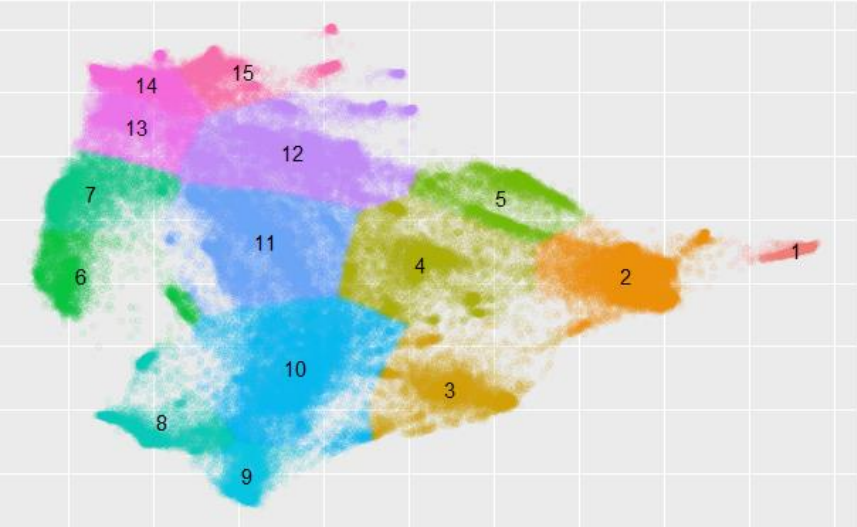
Notfall



Verordnungskosten



Clustering



Cluster	Anteil	Alter	Anteil ♀	Anz. Fälle	Notfall	Verordn.-€
9	3,6%	4,0	50%	4,7	34,5%	70,79 €
8	4,1%	13,2	37%	6,2	26,9%	285,24 €
10	14,9%	27,0	34%	4,9	20,9%	264,44 €
6	3,7%	30,4	50%	6,1	17,9%	146,20 €
1	1,7%	32,0	100%	8,4	28,7%	229,03 €
3	5,8%	32,8	42%	7,3	19,2%	303,59 €
2	9,7%	33,7	99%	8,6	19,0%	145,77 €
4	7,7%	44,7	62%	9,6	19,0%	611,59 €
5	4,7%	49,8	88%	9,9	14,1%	274,85 €
11	12,1%	53,5	50%	9,6	16,1%	463,57 €
12	13,0%	61,1	43%	9,5	13,6%	935,58 €
7	4,7%	70,7	64%	11,9	13,4%	769,21 €
15	4,3%	72,9	66%	10,5	21,5%	2234,36 €
14	5,0%	73,5	39%	11,9	16,6%	1859,78 €
13	4,9%	74,4	43%	14,9	17,6%	1941,36 €
gesamt	*100%*	*45,6*	*54%*	*8,7*	*18,7%*	*654,17 €*

Anwendungsbeispiel: Prädiktion Verordnungskosten

Prospektiv Kosten 2019 aus Diagnosen 2018

Modell	Lin. Regr. R2	LGB Regr. R2	LGB Regr. MAE	LGB Regr. CPM
Alter + Geschlecht	1,0 %	1,1 %	801,08 €	9,4 %
Top-100-Diagn. + Alter + Geschlecht	2,0 %	2,4 %	752,78 €	14,9 %
Pat2Vec 100 + Alter + Geschlecht	7,7 %	**13,7 %**	690,70 €	21,9 %

Ergebnis:
Deutliche Verbesserung bei Verwendung von Pat2Vec+LightGBM

Zusammenfassung und Ausblick

- **Komprimierung aller Diagnosen** von Patient*innen-Profilen
- **Deutliche Verbesserung bei versorgungsrelevanten Prädiktionsmodellen** (im Vergleich zu allgemeinen binarisierten Modellen)
    - Hyperparametertuning notwendig
    - Interpretierbarkeit der Variablen wird schwieriger
- **Datenschutz**: Vektorisierung erschwert Identifikation zusätzlich
- Vortrainiertes Vektorisierungsmodell lässt sich anderen zur Verfügung stellen und auf **viele Fragestellungen der Versorgungsforschung** verwenden

Vielen Dank für Ihre Aufmerksamkeit

Dr. Edgar Steiger
https://www.linkedin.com/in/edgarsteiger

www.zi.de

Zentralinstitut für die kassenärztliche Versorgung in der Bundesrepublik Deutschland
Salzufer 8
10587 Berlin

Tel. +49 30 4005 2450
Fax +49 30 4005 2490
zi@zi.de
@zi_berlin