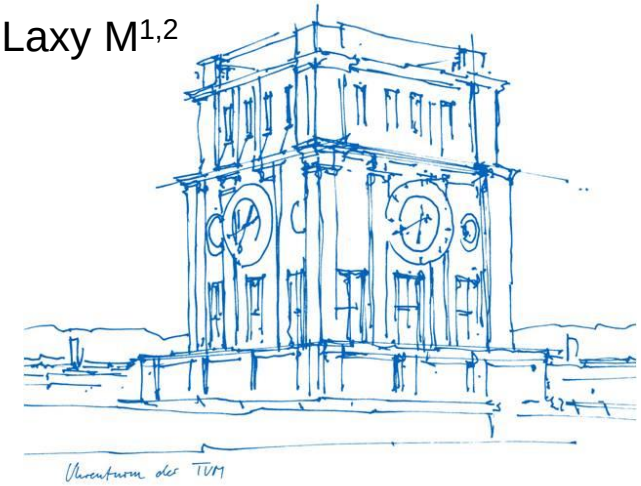
Entwicklung von Machine Learning Algorithmen zur Prädiktion von Diabeteskomplikationen mit GKV-Abrechnungsdaten

Stephan AJ[1,2], Hanselmann M[1], Fan M[1,2], Marsing D[1,3], Laxy M[1,2]

[1]Technische Universität München, München
[2]Helmholtz-Zentrum München, Neuherberg
[3]Ludwig-Maximilians-Universität München, München



Gliederung

- Einleitung
  - Hintergrund
  - Projektziele
  - Geplante Methoden & Status Quo
- Methodische Kernprobleme & Lösungsansätze
  - Prädiktionsendpunkte
  - Beobachtungs- und Prädiktionszeitfenster, Trainings- und Testdaten
  - Feature-Selektion
  - Kriterien für Prädiktionsgüte
- Diskussion

Einleitung
Hintergrund

- „Volkskrankheit“ Diabetes Typ 2 (Prävalenz in Deutschland: ca. 9,5% in 2015[1])
- Bekannte Effekte von Diabeteskomplikationen auf
  - Mortalität & Lebensqualität
  - Kosten im Gesundheitssystem
- Voraussetzung für gezielte Sekundärprävention:

  → Rechtzeitige Identifikation von Patient*innen mit hohem Komplikationsrisiko
- „Nachteile“ vieler bisheriger Prädiktionsmodelle:
  - Klinische Daten als Input: Hoher diagnostischer Aufwand

    → Alternative: GKV-Abrechnungsdaten
  - „Klassische“ parametrische Regressionsmodelle: Begrenzte Anzahl Prädiktoren, linearer Prädiktor

    → Alternative: Methoden des überwachten maschinellen Lernens

[1]Goffrier et al. 2017

Einleitung
Projektziele

1) Entwicklung und Validierung von Prädiktionsmodellen zur Identifikation (und Risikostratifikation) von Diabetespatienten mit hohem Risiko für Komplikationen in GKV-Abrechnungsdaten mit Methoden des maschinellen Lernens
2) Vergleich der Prädiktionsgüte mit einem „klassischen“ logistischen Prädiktionsmodell

Förderer: Deutsches Zentrum für Diabetesforschung 3.0

Einleitung
Geplante Methoden

Datenbasis:

- GKV Abrechnungsdaten Q1 2014 – Q3 2019

Datenumfang:

- Stammdaten, DMP Teilnahme
- Ambulante, stationäre, ambulante stationäre Versorgung & Reha
- Verscheibungen (Arzneimittel, Heil- und Hilfsmittel)

Stichprobe:

- n=912,035 Personen mit Diabetesindikation (Diabetesdiagnose oder DMP Einschreibung)
- Selektionsalgorithmus für gesicherte Diagnosen[2]

→ n= 500,237 (514,654) mit prävalentem Diabetes in 2014 (2015),

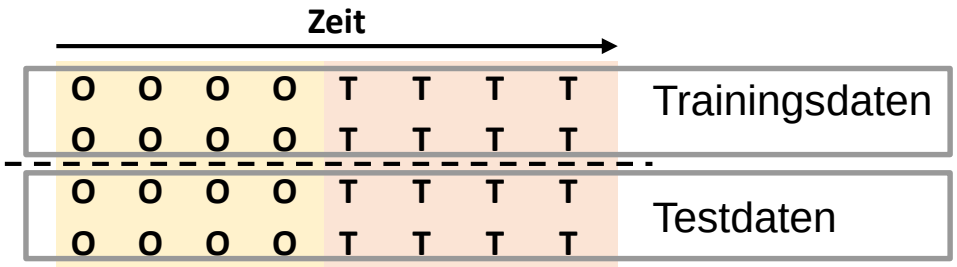
→ 56,97% weiblich, Durchschnittsalter: 71,6 Jahre (19-109)

[2]Kähm et al. 2018

Einleitung
Geplante Methoden & Status Quo

Schritte in der Entwicklung von Prädiktionsmodellen allgemein:

1) Definition von Beobachtungs- und Zielzeitfenstern
2) Aufteilung der Beobachtungen in Trainings- und Testdatensätze
3) Training des Algorithmus (Trainingsdatensatz)
4) Bestimmung der Prädiktionsgüte des Algorithmus basierend auf “neuen” Daten (Testdatensatz, “out-of-sample” & idealerweise “out-of-time”)



Einleitung
Geplante Methoden & Status Quo

Geplante Prädiktionsmodelle:

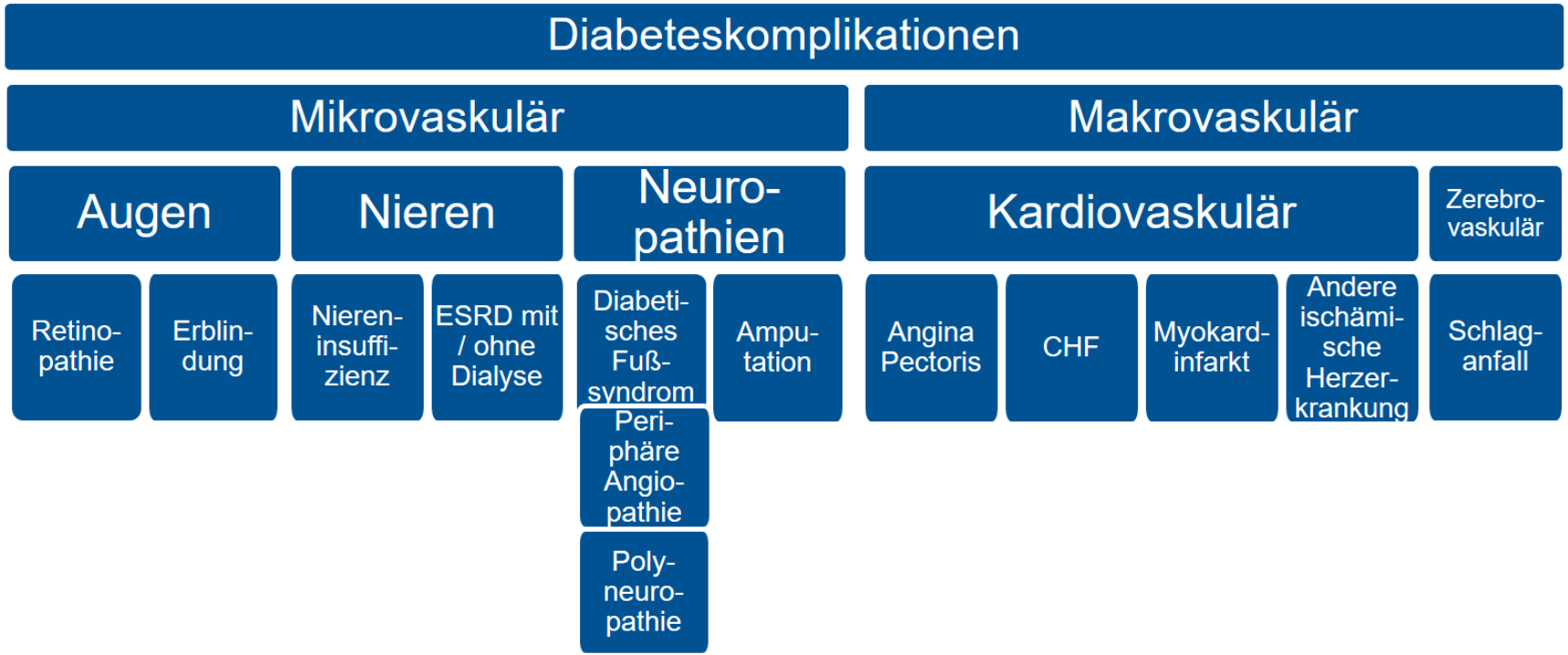
1) Logistische Regression
2) Regularization (e.g. LASSO)
3) Random Forests
4) Gradient Boosting
5) Neural Networks

Status Quo:

- Daten vorhanden, erste Plausibilitätschecks abgeschlossen
- Protokollentwicklung beinahe abgeschlossen

Methodische Kernprobleme & Lösungsansätze

Prädiktionsendpunkte



Methodische Kernprobleme & Lösungsansätze
Prädiktionsendpunkte

→ Austausch mit klinischen Experten:

- Welche Diabeteskomplikation eignet sich am Besten im Hinblick auf
  - Präventabilität
  - Detektierbarkeit
  - Identifizierbarkeit als „inzident“
  - Ökonomische Konsequenzen

Komplikation	Häufigkeit*	Kosten pro Fall**	Häufigkeit x Kosten***	Ranking
Retinopathie	3,8%	99	373	11
Erblindung	0,2%	3741	748	10
Niereninsuffizienz	5,9%	4805	28.351	1
ESRD	0,4%	38.467	15.387	3
Diabetisches Fußsyndrom inkl. Polyneuropathie & Periphäre Angiopathie	5,2%	1.747	9.026	6
Amputation	0,2%	21.628	4.326	9
Myokardinfarkt	0,7%	8.046 – 9.844	3.563 – 5.364 (5.963)	7
Schlaganfall	0,8%	10.523 – 15.622	13.018 – 8.769 (9.251)	5
Angina Pectoris	1,8%	2.769	5.076	8
CHF	4,5%	5865	26.195	2
Andere ischämische Herzerkrankung	0,7%	7.465 – 20.288	4.977 – 13.526 (10.894)	4

* Geschätzt 1-Jahres kumulative Inzidenz nach einem Jahr Komplikationsfreiheit
** Gesamtkosten im ersten Jahr für einen 60- bis 69-jährigen Mann (Kähm et al. 2018)
*** In hypothetischer Kohorte von 100 Diabetikern

→ Entscheidung für Myokardinfakt und Apoplex

[2]Kähm et al. 2018

Methodische Kernprobleme & Lösungsansätze

Beobachtungs- und Prädiktionszeitfenster, Trainings- und Testdaten

J	2014				2015				2016				2017				2018				2019		
Q	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Hauptanalyse: Prävalente Diabetesfälle																							
Sekundäre Analyse 1: Restriktion auf Diabetiker ohne andere Komplikation																							
					O	O	O	O	T	T	T	T	T	T	T	T	T	T	T	T			
Sekundäre Analyse 2: Inzidente Diabetesfälle																							
	R	R	R	R	O	O	O	O	T	T	T	T	T	T	T	T	T	T	T	T			
Optionale Sensitivitätsanalyse 1: Prävalente Diabetesfälle																							
	O	O	O	O	O	O	O	O	T	T	T	T	T	T	T	T	T	T	T	T			
Optionale Sensitivitätsanalyse 2: Prävalente Diabetesfälle																							
					O	O	O	O	T	T	T	T	T	T	T	T							

J: Jahr, Q: Quartal, O: Beobachtungszeit, T: Prädiktionszeit, R: Run-in Periode.

Cave: Testdaten sind „out-of-sample“, aber nicht „out-of-time“

Methodische Kernprobleme & Lösungsansätze
Feature-Selektion

- Möglichkeit der Aufnahme einer hohen Anzahl potenzieller „Prädiktoren“
- Abwägung:

„Garbage In, Garbage Out“ ⟷ Potenzial für neue Erkenntnisse

- Ansätze:

1) Literaturgestützte Auswahl[3,4]:

  – andere Prädiktionsmodelle zu Diabetes
  – andere ML Modelle für Sekundärdaten

  → ca. 165 Potenzielle Features, davon
    - ~ 14 sozio-demographisch
    - ~ 10 inanspruchnahmebezogen
    - ~ 8 kostenbezogen
    - ~ 42 Komorbiditätsindizes & Einzeldimensionen
    - ~ 20 weitere Komorbiditäten
    - ~ 57 Medikamentenkategorien
    - ~ 14 andere Diabeteskomplikationen

[3]Cichosz et al. 2016, [4]Ndjaboue et al. 2021

Methodische Kernprobleme & Lösungsansätze
Feature-Selektion

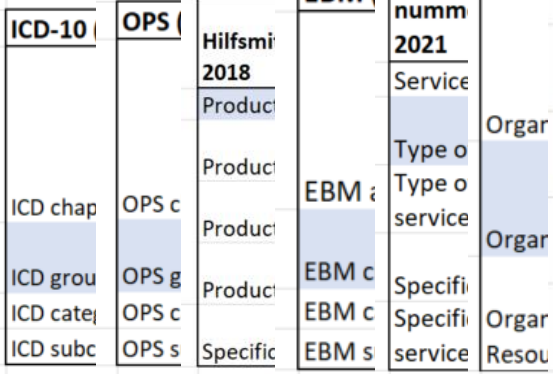
- Möglichkeit der Aufnahme einer hohen Anzahl potenzieller „Prädiktoren“
- Abwägung:

„Garbage In, Garbage Out“ ⟷ Potenzial für neue Erkenntnisse

Codieraufwand & praktische Implementierbarkeit ⟷ Aussagekraft

- Ansätze:

2) „Agnostisch“

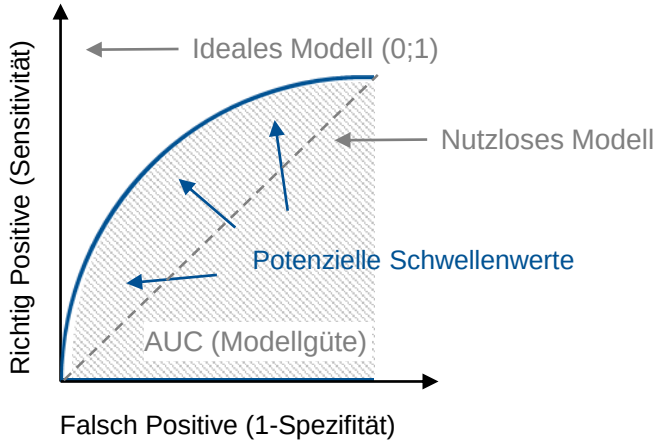


ATC (2021)	Number	Description
ATC level 1	14	anatomic group, first position letter
Up to ATC level 2	95	therapeutic main group, second and third position digits
Up to ATC level 3	267	therapeutic/pharmacological subgroup, fourth position letter
Up to ATC level 4	889	chemical/therapeutic/pharmacological subgroup, fifth position letter
Up to ATC level 5	5067	chemical substance subgroup, sixth and seventh position digits

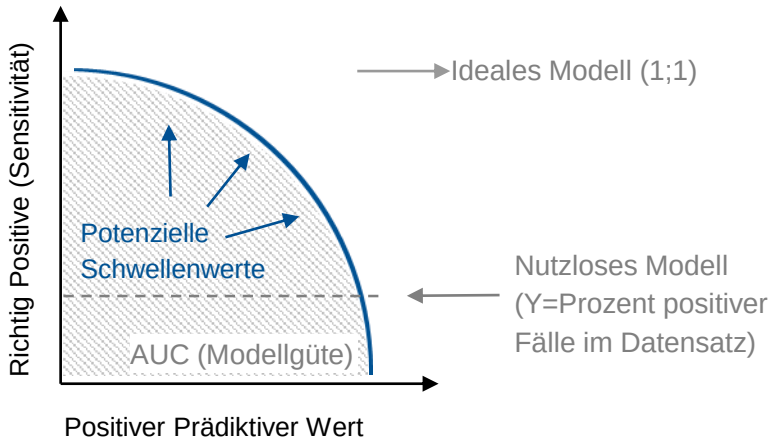
Methodische Kernprobleme & Lösungsansätze
Kriterien für Prädiktionsgüte

- Diskrimination: Auswahl des optimierten Schwellenwerts basierend auf

a) C-Statistik (Concordance, AUROC) anhand des Youden Index
→ balanciert Sensitivität und Spezifität
(unabhängig von der Prävalenz des Outcomes)



b) Precision-Recall Kurven anhand des F1 Scores → gewichteter Durchschnitt von Präzision (PPV) und Recall (Sensitivität)
(Fokus auf Positive, Negative nicht betrachtet)



→ Precision-Recall aufgrund der geringen erwarteten Outcome-Inzidenzen (unbalanciertes Outcome)

Diskussion

Es existieren multiple Lösungsansätze mit jeweils Vor- und Nachteilen für die methodischen Herausforderungen in Bezug auf

- Prädiktionsendpunkte
- Beobachtungs- und Prädiktionszeitfenster, Trainings- und Testdaten
- Feature-Selektion
- Kriterien für Prädiktionsgüte

→ Abwägung im Hinblick auf technische Machbarkeit, Ressourcenintensität, und Konsequenzen für die praktische Nützlichkeit des resultierenden Algorithmus.

→ Ausblick: Besseres Verständnis des Potenzials von „Machine Learning“ Algorithmen in GKV-Abrechnungsdaten & verbundenen methodischen Herausforderungen.

Vielen Dank für Ihre Aufmerksamkeit!

Referenzen

1. Goffrier, Benjamin, Mandy Schulz, and Jörg Bätzing-Feigenbaum. "Administrative Prävalenzen und Inzidenzen des Diabetes mellitus von 2009 bis 2015." *Zentralinstitut für die kassenärztliche Versorgung in Deutschland (Zi). Versorgungsatlas-Bericht* 17/03 (2017)

2. Kähm, Katharina, et al. "Health care costs associated with incident complications in patients with type 2 diabetes in Germany." *Diabetes Care* 41.5 (2018): 971-978.

3. Cichosz, Simon Lebech, Mette Dencker Johansen, and Ole Hejlesen. "Toward big data analytics: review of predictive models in management of diabetes and its complications." *Journal of diabetes science and technology* 10.1 (2016): 27-34.

4. Ndjaboue, Ruth, et al. "A scoping review of the predictive models of diabetes complications." (2021).

Professorship for Public Health & Prevention

Technische Universität München
Georg-Brauchle-Ring 60/62
80992 München

https://www.sg.tum.de/php/startseite/

@TUMPublicHealth

Anna-Janina Stephan, MPH, PhD

Room: L-201
Tel: +49 (0) 89 / 289 24984
anna-janina.stephan@tum.de

MNGHC Team

Prof. Dr. Michael Laxy	michael.laxy@tum.de
Dr. Michael Hanselmann	michael.hanselmann@tum.de
Min Fan	min.fan@tum.de