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

Hintergrund: Diabeteskomplikationen beeinträchtigen die Lebensqualität der Betroffenen und verursachen hohe Kosten im Gesundheitssystem. Die rechtzeitige Identifikation von Patient*innen mit hohem Risiko ist Voraussetzung für gezielte Maßnahmen der Sekundärprävention. Moderne Methoden des maschinellen Lernens bieten die Möglichkeit, anhand bestehender Muster in Sekundärdaten Diabetespatient*innen mit hohem Komplikationsrisiko zu identifizieren.

Ziel: Ziel dieser Studie ist es, mithilfe von modernen Methoden des maschinellen Lernens (Random Forests, Gradient Boosting, neuronale Netze) auf Basis von GKV-Abrechnungsdaten Algorithmen zur Prädiktion von Diabeteskomplikationen und zur Risikostratifikation zu entwickeln, zu validieren und deren Prädiktionsgüte mit klassischen logistischen Modellen zu vergleichen.

Methodische Kernprobleme: Kernprobleme umfassen die 1) Selektion geeigneter Diabeteskomplikationen als Prädiktionsendpunkte, 2) Auswahl geeigneter „Features“ (Prädiktoren) zum Training der Algorithmen, 3) Definition von Beobachtungs- und Prädiktionszeitfenstern, 4) Aufteilung auf Trainings- und Testdatensätze sowie 5) Auswahl von Kriterien zur Beurteilung der Prädiktionsgüte.

Lösungsansätze: Die Auswahl der Endpunkte wird in Kooperation mit klinischen Experten anhand der Kriterien Präventabilität, Detektierbarkeit, ökonomische Konsequenzen, und Definierbarkeit als „inzident“ eruiert. Die Feature-Selektion stützt sich auf bestehende Diabetes-Prädiktionsmodelle, publizierte „Machine-Learning“ Modelle mit Kassendaten in anderen Indikationen und eine expertengestützte Auswahl zusätzlich potenziell aussagekräftiger Abrechnungscodes. Sensitivitätsanalysen sollen den Einfluss der Beobachtungs- und Prädiktionszeitfenster beleuchten. Die Aufteilung auf Trainings- und Testdatensätze erfolgt anhand etablierter Kriterien. Es werden je mehrere Kennzahlen für Diskrimination und Kalibrierung, aber auch eine Entscheidungskurven-Analyse hinsichtlich des klinischen Nutzens berichtet.

Diskussion: Für die genannten methodischen Herausforderungen existieren multiple Lösungsansätze mit jeweils Vor- und Nachteilen. Diese gilt es im Hinblick auf technische Machbarkeit, Ressourcenintensität, und Konsequenzen für die praktische Nützlichkeit des resultierenden Algorithmus abzuwägen.

Schlussfolgerungen: Dieses Projekt soll zu einem besseren Verständnis des Potenzials von „Machine Learning“ Algorithmen in GKV-Abrechnungsdaten sowie der damit verbundenen methodischen Herausforderungen beitragen.

Korrespondenzadresse der Erstautorin: Dr. Anna-Janina Stephan, TU München, Fakultät für Sport- und Gesundheitswissenschaften, Professur für Public Health und Prävention, Georg-Brauchle-Ring 60/62, 80992 München