Kaya Miah
Dr. sc. hum.

Model selection in the framework of multi-state models

Fach/Einrichtung: Medizinische Biometrie, Deutsches Krebsforschungszentrum (DKFZ)
Doktormutter: Prof. Dr. Annette Kopp-Schneider

In der medizinischen Forschung werden in Prognosemodellen überwiegend zusammengesetzte Endpunkte wie das progressions- oder ereignisfreie Überleben verwendet. Diese Überlebenszeitendpunkte für die Zeit bis zum Auftreten des ersten Ereignisses lassen jedoch wichtige Aspekte des individuellen Krankheitsverlaufs und der Therapie unberücksichtigt. Mehrstadienmodelle sind ein nützliches methodisches Konzept, um Effekte von prognostischen Faktoren und Behandlungen auf den Ereignisverlauf eines Patienten zu schätzen und die Risiken für das Auftreten verschiedener Ereignisse zu separieren. Sie erweitern die Analyse konkurrierender Risiken für Endpunkte wie die Zeit bis zum Fortschreiten der Erkrankung, Rezidiv, Remission oder Tod, indem sie die Abfolge konsekutiver Zustände modellieren.

Diese Arbeit wurde durch eine Anwendung auf den Krankheitsverlauf der akuten myeloischen Leukämie (AML) motiviert. Um zu beurteilen, wie die Wahrscheinlichkeit, von einem Zustand in einen anderen zu wechseln, von Kovariablen abhängt, können proportionale Hazard-Regressionsmodelle im Mehrstadienkontext verwendet werden. Im Rahmen der Präzisionsmedizin mit hochdimensionaler Information in Form von molekularen Biomarkern ist eine solche holistische Analyse eines Mehrstadienmodells von wesentlichem Interesse. Für die motivierende AML Anwendung wurde der Einfluss von Biomarkern in Form von Genmutationen zusammen mit etablierten klinischen Prädiktoren auf die Übergänge eines 9-Stadienmodells untersucht. Dabei sind wirksame Strategien zur Variablenselektion für Mehrstadienmodelle basierend auf hochdimensionalen Daten erforderlich, um ein schwach besetztes Modell zu erhalten und eine Überanpassung zu vermeiden. Solche datengetriebenen Modellbildungsstrategien tragen zu einem tieferen Verständnis des individuellen Krankheitsverlaufs und seiner Therapiekonzepte sowie zu verbesserten Prognosen bei.

In dieser Arbeit wurden *Fused Sparse-Group Lasso* (FSGL) penalisierte Mehrstadienmodelle für die datengetriebene Variablenselektion vorgeschlagen, um pathogene Krankheitsprozesse unter Einbeziehung klinischer und molekularer Daten genauer zu erfassen. Ziel war es, ein schwach besetztes Modell auf Grundlage hochdimensionaler Daten mittels erweiterter Regularisierungsverfahren zu selektieren. Der *Alternating Direction Method of Multipliers* (ADMM) Algorithmus wurde für FSGL penalisierte Mehrstadienmodelle adaptiert. Dieser *FSGLmstate*-Algorithmus kombiniert die Penalisierungskonzepte der allgemeinen Variablenselektion, paarweisen Differenzen von Kovariableneffekten sowie der Gruppierung von Übergängen. Auf diese Weise erreichen FSGL penalisierte Mehrstadienmodelle eine dimensionsreduzierte Modellbildung unter Einbeziehung von a-priori Informationen über die Kovariablen- und Übergangsstruktur. Des Weiteren kann der ADMM-Algorithmus aufgrund der Zerlegbarkeit des Optimierungsproblems hochdimensionale Problemstellungen bewältigen. Mittels einer *Proof-of-Concept*-Simulationsstudie wurde die Regularisierungsleistung des FSGLmstate-Algorithmus

evaluiert, um ein Modell zu selektieren, welches nur relevante übergangsspezifische Effekte sowie ähnliche übergangsübergreifende Effekte enthält. Im Gegensatz zu nicht-penalisierten und globalen *Lasso*-penalisierten Schätzungen identifiziert FSGLmstate Ähnlichkeits- und Gruppierungsstrukturen in Abhängigkeit von der Wahl der Penalisierungsparameter. Die Anwendung auf eine klinische Phase III Studie für die AML veranschaulicht den Nutzen eines FSGL penalisierten Mehrstadienmodells zur Reduzierung der Modellkomplexität bei gleichzeitiger Berücksichtigung von klinischen und molekularen Daten. Durch die Anwendung des FSGLmstate-Ansatzes wird im Gegensatz zu unpenalisierten Mehrstadienmodellen eine Überanpassung vermieden.

Eine Limitation der aktuellen Arbeit besteht darin, dass zeitabhängige Kovariablen, wie beispielsweise die allogene Stammzelltransplantation, sowie zeitabhängige Effekte noch nicht berücksichtigt werden. Zudem muss die Inferenz nach Modellselektion mittels Penalisierung weiter untersucht werden. Aufgrund der Rechenintensität würde der Algorithmus von einer Erhöhung der Recheneffizienz profitieren, um die Leistungsfähigkeit in sehr hohen Dimensionen effizient zu verbessern. Weiterhin sind umfangreiche Simulationsstudien für empirische Methodenvergleiche erforderlich, um die Leistungsfähigkeit des entwickelten Variablenselektionsverfahrens in einem breiten Spektrum von Szenarien zu evaluieren.

Zusammenfassend ist festzuhalten, dass die vorliegende Dissertation eine umfassende Untersuchung von Modellselektionsverfahren für komplexe Mehrstadienmodelle darlegt. Die Arbeit schlägt einen erweiterten Penalisierungsansatz als datengetriebene Variablenselektionsstrategie vor, welche die Konzepte von allgemeiner Sparsamkeit, Ähnlichkeit von Regressionseffekten und übergangsweiser Gruppierung kombiniert, einhergehend mit einem flexiblen Algorithmus (FSGLmstate) sowie zugehöriger Softwareimplementierung.