Analyse von Überlebenszeiten bei hochdimensionalen Daten

Messow CM, Victor A, Hommel G, Blettner M
Institut für Medizinische Biometrie, Epidemiologie und Informatik, Universität Mainz, Deutschland
messow@imbei.uni-mainz.de

Einleitung und Fragestellung Die Anwendung von Standardmethoden zur Überlebenszeitanalyse wird bei Untersuchung vieler möglicher Einflussfaktoren bei einer realistischen Patientenzahl problematisch. Dies ist beispielsweise der Fall, wenn pro Patient viele Laborparameter gemessen werden, oder aber bei der Analyse von Genexpressionsdaten, wo häufig Tausende von Messwerten pro Patient vorhanden sind. Wollte man alle Variablen in ein Modell aufnehmen, bräuchte man eine nur schwer zu bewältigende Anzahl von Patienten. Gerade bei teureren Verfahren wie zum Beispiel bei der Verwendung von Microarrays ist diese große Anzahl schwer oder unmöglich zu erreichen. Ein weiteres Problem, das in der Praxis häufig auftritt, ist, dass die Variablen untereinander korreliert sind, wie das auch bei den oben genannten Beispielen von Laborparametern oder Genexpressionsdaten der Fall ist.
Gerade aus Sicht der Patienten ist nicht nur das Entdecken von prognostischen Faktoren, sondern auch deren Verwendung zur Risikoquantifizierung wichtig. Es stellt sich also die Frage, mit welcher Methodik man unter diesen Voraussetzungen eine sinnvolle Analyse der Überlebenszeiten durchführen kann, um daraus eine möglichst gute Risikoabschätzung für Individuen zu erhalten. Ziel ist es, einen Risikoscore zu bilden, der eine bessere Prognose ermöglicht. Dabei ist entscheidend, dass ausreichend Information aus den Daten verwendet wird, um einen verlässlichen Score zu erhalten, aber gleichzeitig nicht zu viel Information verwendet werden darf, um einen Overfit an die Daten zu vermeiden. Das Ergebnis sollte sich auf andere Patientenkollektive verallgemeinern lassen.
Zu diesem Thema sind bereits diverse Vorschläge publiziert. Ziel dieser Arbeit ist es, anhand verschiedener Datenbeispiele zu überprüfen, wie gut diese Methoden in der Praxis tatsächlich funktionieren.
Material und Methoden Wang et al. [1] haben eine Methode beschrieben, die versucht, mit Hilfe von univariaten Cox-Regression und Auswahl über ROC-(Receiver-Operating-Characteristic) -Kurven einen Risikoscore zu bilden. Um den Score robuster zu gestalten, kommt zusätzlich bei der Modellwahl Bootstrap zum Einsatz.
Bei ihrem Vorgehen stellt sich das Problem der Dichotomisierung der Information über das Überleben, da für die ROC-Kurven eine dichotome Zielvariable benötigt wird. Es muss also ein Zeitpunkt gewählt werden, zu dem betrachtet wird, ob ein Patient bereits ein Ereignis hatte oder nicht. Wählt man den Zeitpunkt relativ spät, sind bis dahin bei vielen Patienten Ereignisse aufgetreten, allerdings könnten dann auch viele Patienten vor dem Zeitpunkt zensiert sein und deshalb nicht miteinbezogen werden. Wählt man den Zeitpunkt früh, verliert man nur wenige Patienten aufgrund von Zensierungen, allerdings sind auch noch nicht so viele Ereignisse eingetreten. Das ist vor allem bei selteneren Ereignissen problematisch.
Als Risikoscore wird der lineare Prädiktor der Cox-Regression verwendet. Bei der Wahl des Modells, das letzten Endes verwendet werden soll, wird als Kriterium für die Modellgüte die Fläche unter der ROC-Kurve (AUC) des Scores aus der Cox-Regression verwendet, mit der dichotomisierten Überlebensinformation als Zielgröße.
Der lineare Prädiktor aus dem so gewählten Modell wird Risikoscore für die einzelnen Patienten verwendet. Dann wird anhand einer ROC-Kurve eine Schwelle für den Score festgelegt, bei dem man die gewünschte Sensitivität bei maximaler Spezifität erhält. An dieser Schwelle werden die Patienten in eine Gruppe mit guter und eine Gruppe mit schlechter Prognose eingeteilt.
Der so erhaltene Score sowie die festgelegte Schwelle müssen anschließend auf einem Testdatensatz validiert werden.
Dieses Verfahren wurde zum einen auf einen realen Datensatz angewendet, zum anderen auf simulierte Daten. Wir simulierten Datensätze mit 1000 möglichen Einflussfaktoren bei 200 Beobachtungen. Die Einflussfaktoren sind zum Teil unkorreliert, zum Teil über verschiedene Korrelationsstrukturen voneinander abhängig. Die Überlebenszeiten wurden exponential verteilt erzeugt. Es wird jeweils ein Trainings- und ein Testdatensatz erstellt.
Ergebnisse Zunächst fällt auf, dass sich die optimale Anzahl der ins Modell aufzunehmenden Variablen nicht eindeutig über die Betrachtung der AUC in Abhängigkeit von der Anzahl der Variablen im Modell erkennen lässt. Die AUC zeigt mehrere Plateaus. Mehrere dieser möglichen Modelle werden weiterbetrachtet. Auch bei der anschließenden Wahl der Schwelle für gute und schlechte Prognose über die ROC-Kurve des Scores kommen wieder mehrere Schwellenwerte in Frage. Wieder werden mehrere mögliche Schwellen angenommen. Bei der Validierung auf dem Testdatensatz zeigt der Score nicht immer gute prognostische Fähigkeiten, und die Einteilung der Testpatienten in gute und schlechte Prognose ist nicht unbedingt gut.
Diskussion Die ersten Ergebnisse deuten an, dass die Methode sinnvolle Lösungsansätze zu der beschriebenen Problematik birgt. Die Verwendung von Bootstrap in der Modellbildung scheint ein sinnvoller Ansatz zu sein. Allerdings hat die Methode einige Nachteile. Problematisch ist, dass die Information über das Überleben an einigen Stellen dichotomisiert werden muss. Das ist nicht nur ein Informationsverlust, sondern bietet auch die Möglichkeit, durch die Wahl des Zeitpunktes das Ergebnis in die eine oder andere Richtung zu beeinflussen. Ebenso muss bei der Modellbildung eine Entscheidung über eine ausreichend große AUC und bei der Festlegung der Schwelle über die gewünschte Sensitivität getroffen werden, die relativ subjektiv ist. Das ermöglicht zwar zum einen die Möglichkeit, das Verfahren möglichst gut an die vorliegende Situation anzupassen, zum anderen können die Ergebnisse so aber auch wieder zu einem Overfit verleiten. Die Methode bietet daher keinen Schutz vor Overfitting, das kann nur durch sinnvolle Entscheidungen durch den Anwender erreicht werden.
Ein weiterer Nachteil der Methode ist, dass die Abhängigkeiten der Variablen untereinander nicht berücksichtigt werden, da sie aufgrund ihres p-Wertes in der univariaten Cox-Regression ins Modell aufgenommen werden oder nicht. Messen zwei Variablen im Wesentlichen dasselbe, hängen sie auch etwa gleich stark vom Überleben ab und somit ist es relativ wahrscheinlich, dass sich beide im Modell wiederfinden.

Literatur

[1] Wang Y, Klijn JGM, Zhang Y, Sieuwerts AM, Look MP, Yang F, Talantov D, Timmermans M, Meijer-van Gelder ME, Yu J, Jatkoe T, Berns EMJJ, Atkins D, Foekens JA. Gene expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 2005; vol. 365, Issue 9460: 671-9