Analyse von Überlebenszeiten bei hochdimensionalen Daten

C.-M. Messow, A. Victor, G. Hommel, M. Blettner

Institut für Medizinische Biometrie, Epidemiologie und Informatik, Uni Mainz

11.09.2006

Grundlegende Probleme

- viele Messungen pro Individuum, z.B. viele Laborparameter, Microarrays
  $\Rightarrow p >> N$
- Variablen sind untereinander korreliert

Dann: Standardmethoden der Survivalanalyse nicht anwendbar

Yixin Wang, Jan G M Klijn, Yi Zhang, Anieta M Sieuwerts Maxime P Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E Meijer-van Gelder, Jack Yu, Tim Jatkoe, Els M J J Berns, David Atkins, John A Foekens (2005):
Gene expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer

Ziel:
Basierend auf Genexpressionsdaten soll ein Entscheidungsverfahren zur Vorhersage des Überlebens bei Patientinnen mit Brustkrebs entwickelt werden.

Methode

- Aufteilung der Daten in Lern- und Testdatensatz
- Bestimmung des Stichprobenumfangs des Lerndatensatzes
- Genselektion mit Cox-Regression und Bootstrap im Lerndatensatz
- Bildung des Relapse Score im Lerndatensatz
- Validierung auf Testdatensatz

Methode

- Aufteilung der Daten in Lern- und Testdatensatz
- Bestimmung des Stichprobenumfangs des Lerndatensatzes
- Genselektion mit Cox-Regression und Bootstrap im Lerndatensatz
- Bildung des Relapse Score im Lerndatensatz
- Validierung auf Testdatensatz

Auswahl der Gene

Univariate Cox-Regression für jedes Gen mit Bootstrap:

- 400 Samples aus den Patienten aus dem Training Set ziehen (mit Zurücklegen, gleicher Stichprobenumfang)
- pro Sample und Gen eine Cox-Regression
- pro Gen wird Score berechnet als um 5% gestutztes Mittel der 400 p-Werte ($\rightarrow$ Bootstrap-p-Wert)
- Gene nach den Bootstrap-p-Werten ordnen

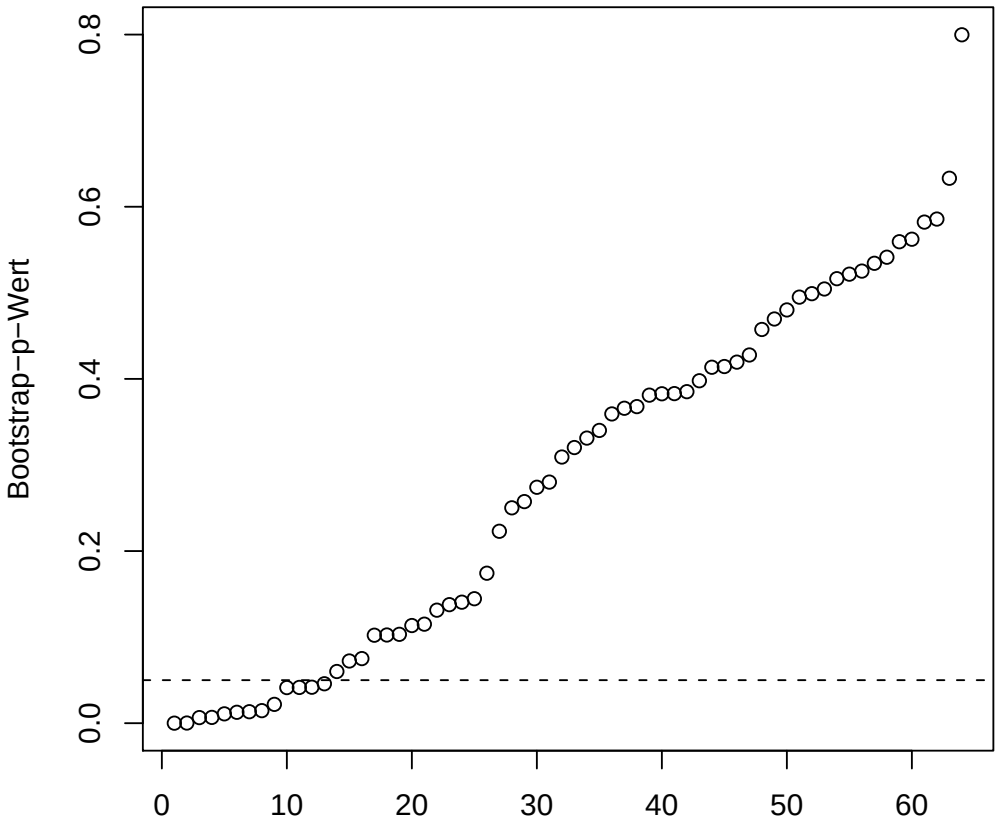
Bildung des Relapse Score

- Überleben dichotomisieren (Überleben zum Zeitpunkt t)
- Logistische Regression, nacheinander immer eine Variable mehr ins Modell bis die Fläche unter der zugehörigen ROC-Kurve (AUC) maximal (bzw. ausreichend groß) ist
- Bildung des Scores als Summe über die Expression der gewählten Gene gewichtet mit den zugehörigen Cox-Regressionskoeffizienten ($x_i'\beta$)
- Schwelle für gute / schlechte Prognose für Relapse Score bei 100% Sensitivität und maximaler Spezifität

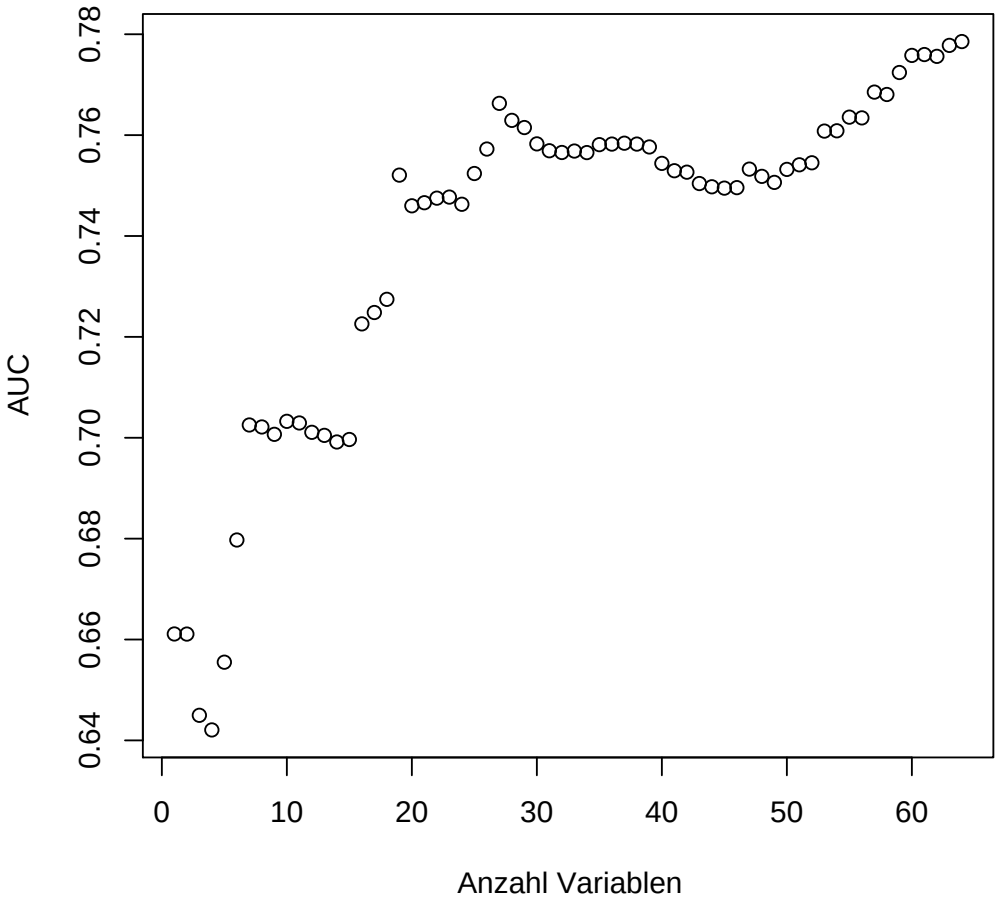
Daten zu Risikofaktoren bei Koronarer Herzkrankheit

- Untersuchung des Einflusses diverser Risikofaktoren auf Überleben bei KHK
- Endpunkt: Kardiovaskulärer Tod oder Myokardinfarkt
- Potentielle Einflussgrößen sind Alter, BMI und 62 Laborparameter
- fehlende Werte werden entsprechend der Variable normalverteilte Zufallszahlen ersetzt
- 1956 Patienten, davon $\frac{1}{3}$ Testdatensatz
- insgesamt 122 Ereignisse, davon 44 im Testdatensatz

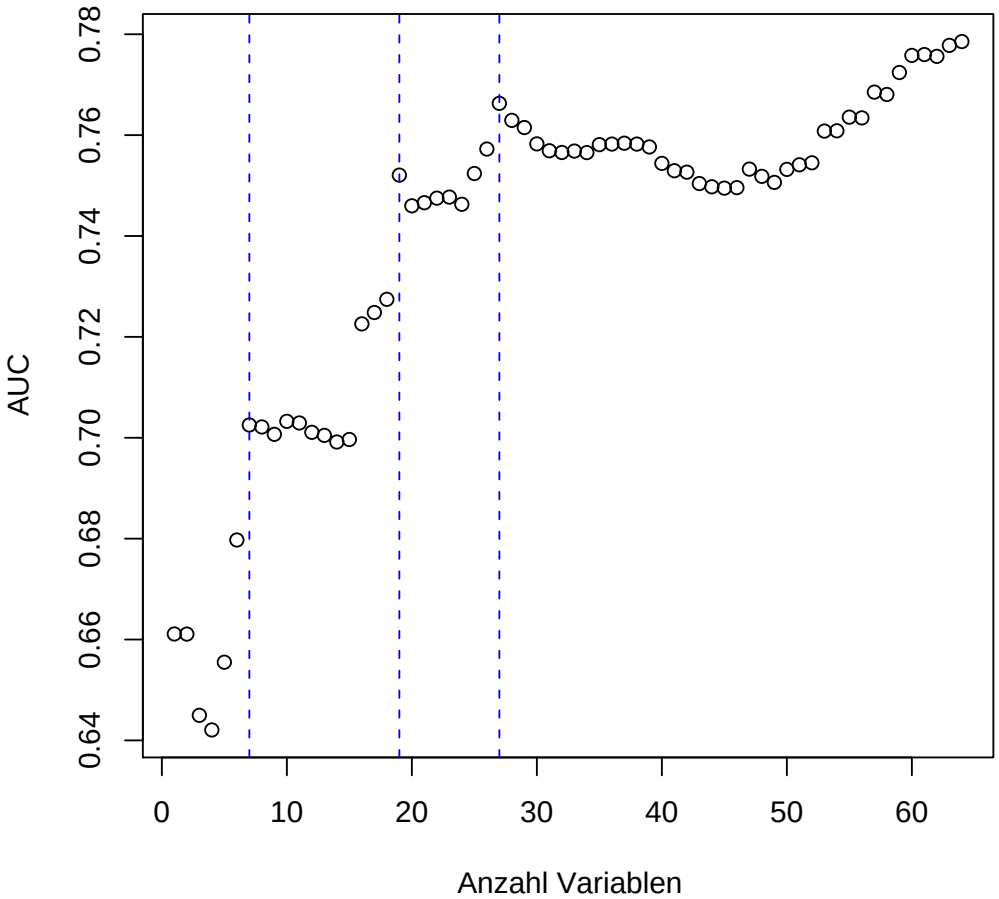
Ergebnisse



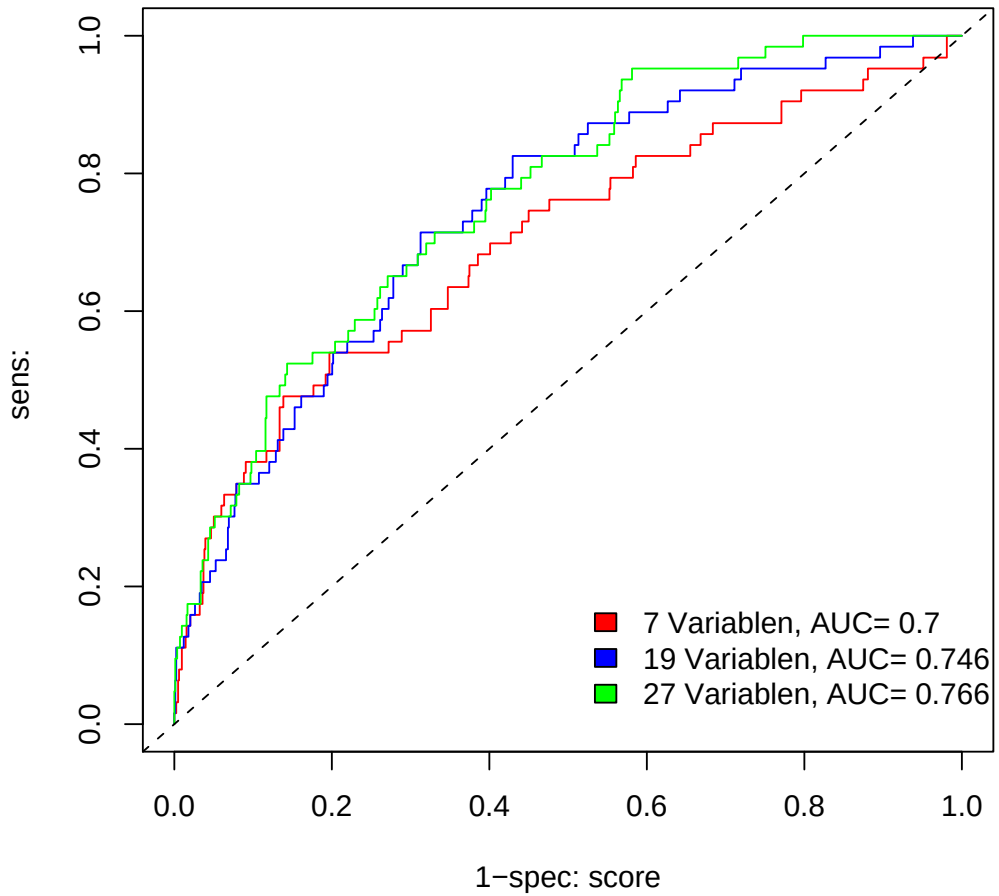
AUCs der ROC-Kurven nach Anzahl der Variablen



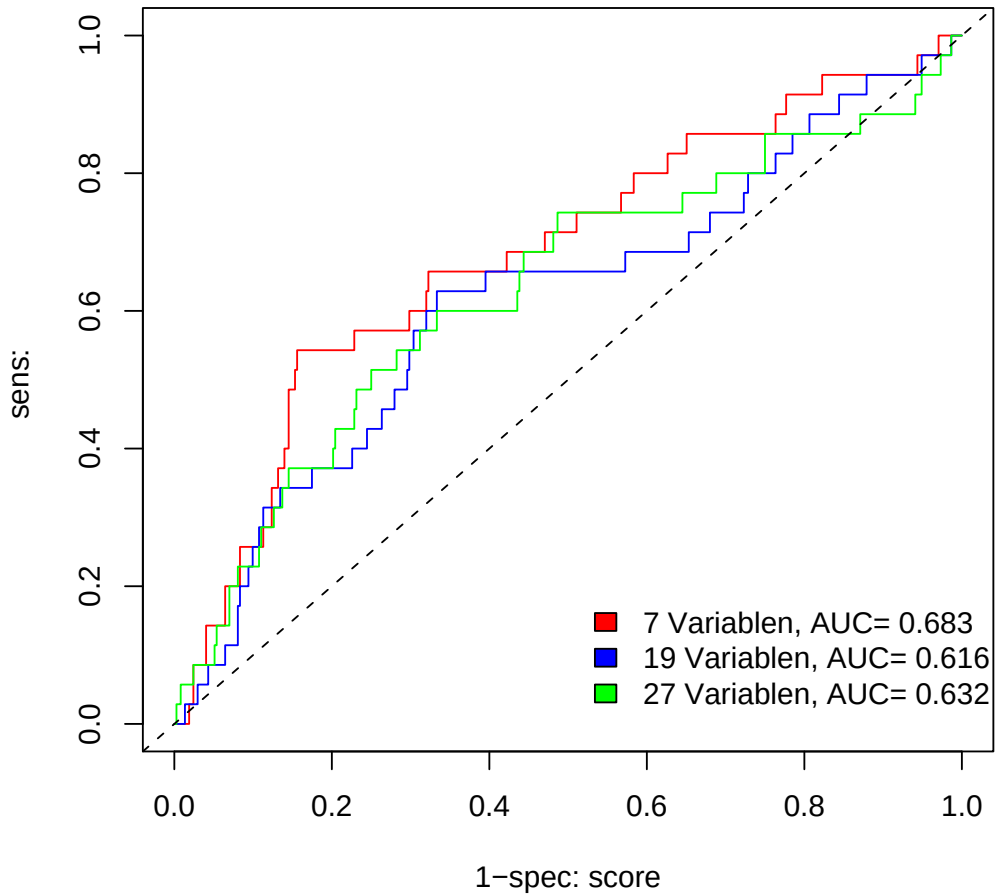
AUCs der ROC-Kurven nach Anzahl der Variablen



ROC-Kurven der Relapse-Scores im Lerndatensatz



ROC-Kurven der Scores auf dem Testdatensatz



Simulationen

- je 1000 normalverteilte Einflussgrößen, abhängig oder unabhängig voneinander
- je 200 Individuen
- β wird vorgegeben, unterschiedlicher Anteil an Variablen mit Einfluss
- Überlebenszeiten exponentialverteilt, abhängig von $X\beta$ (Bender et al., 2005)
- Beobachtungszeiten exponentialverteilt
- 25% / 50% Ereignisse
- je 100 Simulationen
- Testdaten gleiche Struktur

Bildung des Scores

- feste Anzahl Variablen (15 / 10)
- AUC > 0.8
- AUC > 0.9
- AUC > 0.8 und Plateau
- AUC=1

Einfluss der Variablen

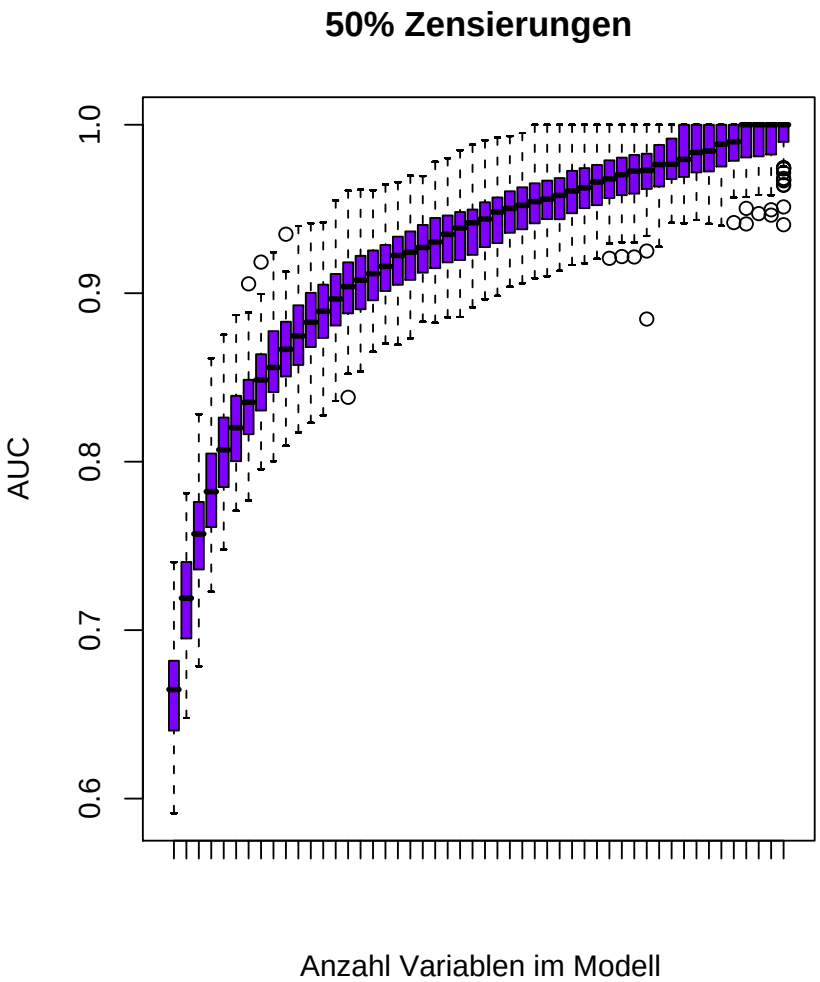
Variante 1:

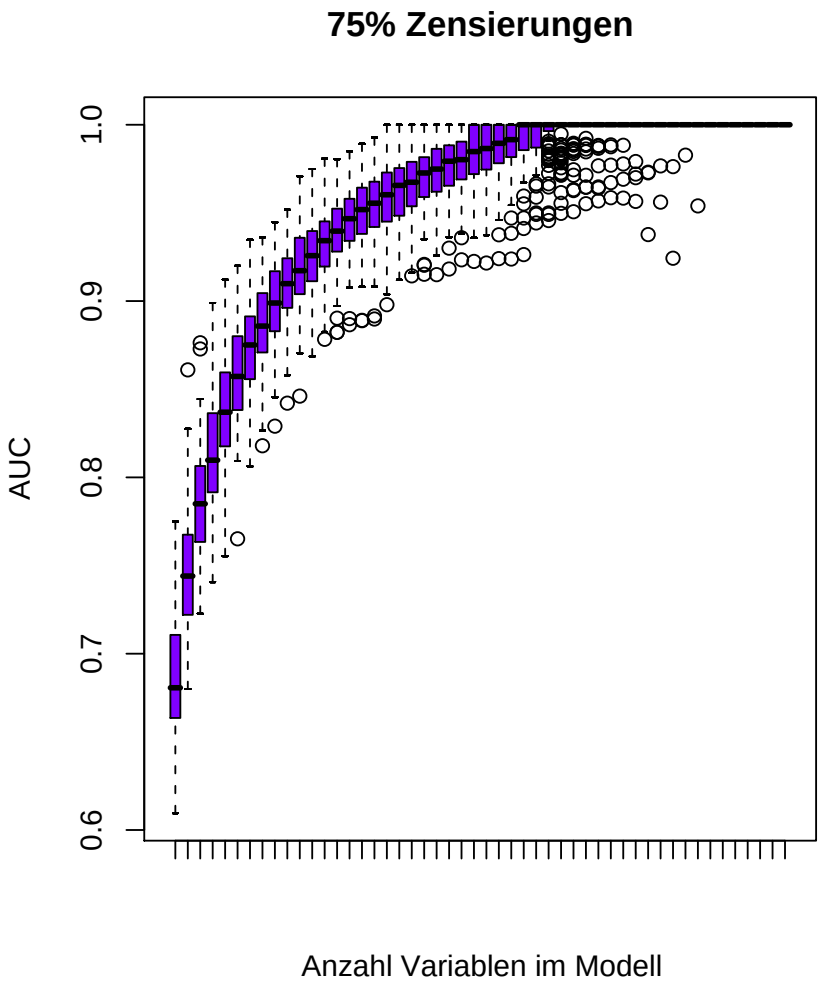
$\beta_1, \ldots, \beta_{10} = \log(5)$
$\beta_{11}, \ldots, \beta_{20} = \log(3)$
$\beta_{21}, \ldots, \beta_{30} = \log(2)$
$\beta_{31}, \ldots, \beta_{50} = \log(1.5)$
$\beta_{51}, \ldots, \beta_{1000} = 0$

Variante 2:

$\beta_1, \ldots, \beta_{10} = \log(1.5)$
$\beta_{11}, \ldots, \beta_{20} = \log(1.2)$
$\beta_{21}, \ldots, \beta_{30} = \log(1.1)$
$\beta_{31}, \ldots, \beta_{1000} = 0$

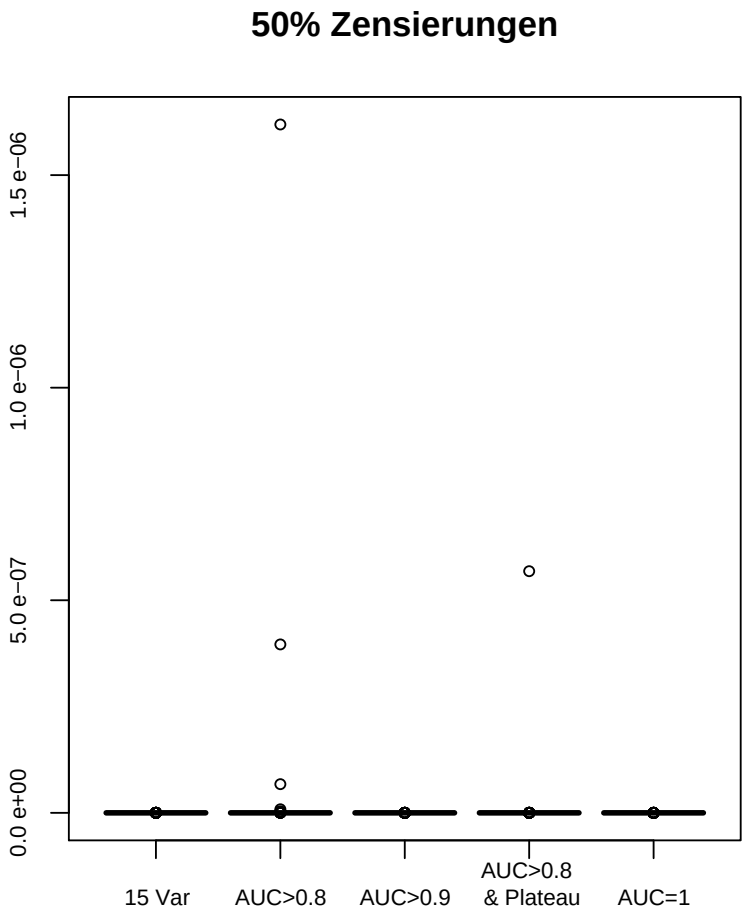
Unabhängige Einflussgrößen

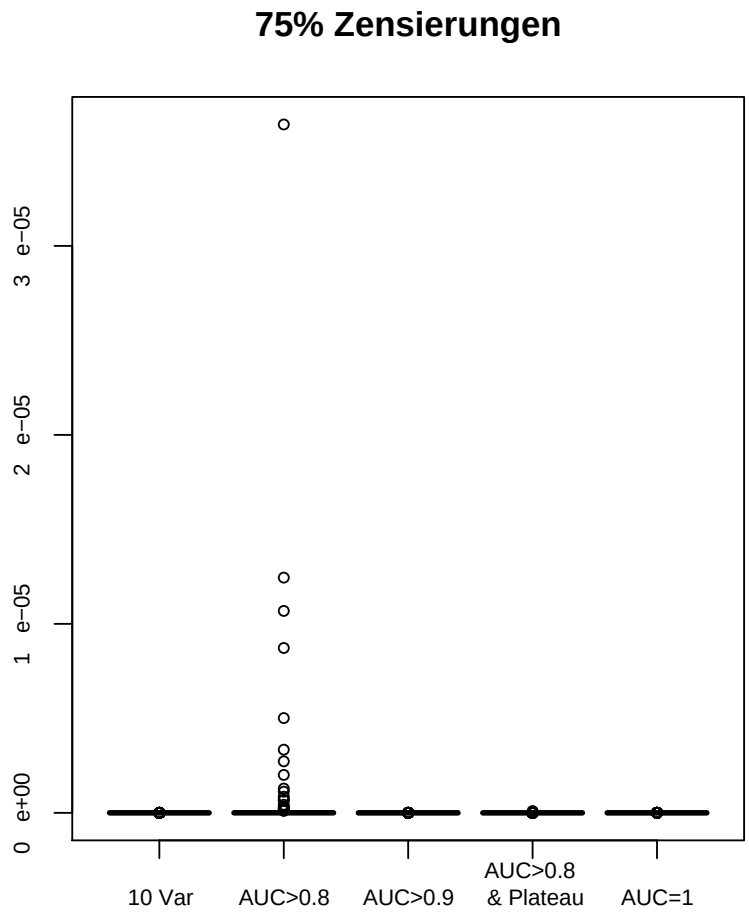




Unabhängige Einflussgrößen

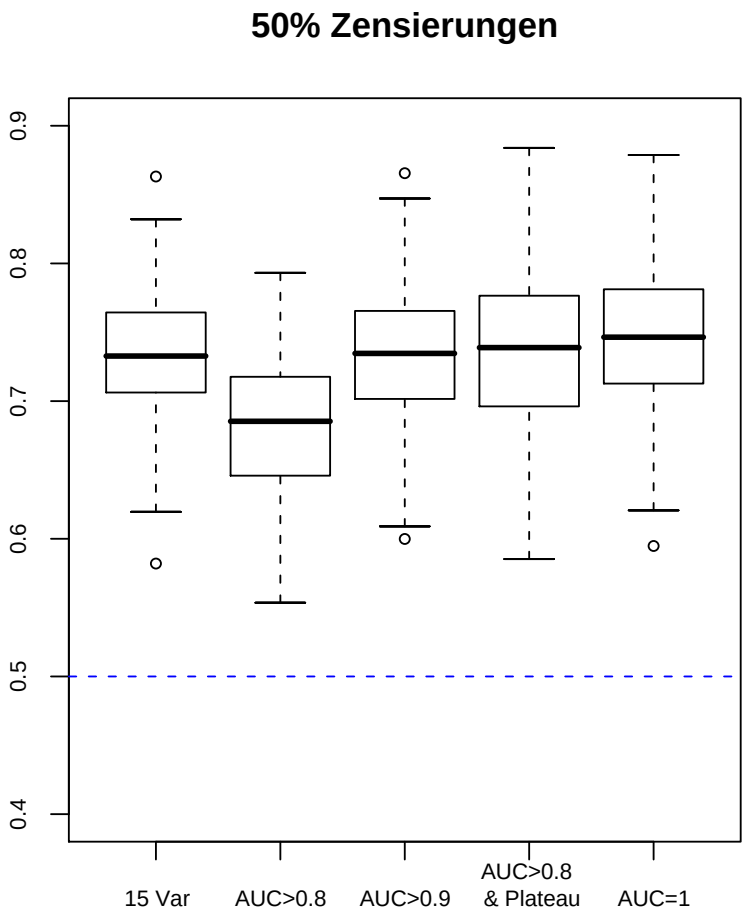
p-Werte für gute vs. schlechte Prognose, 80% Sensitivität, Lerndaten

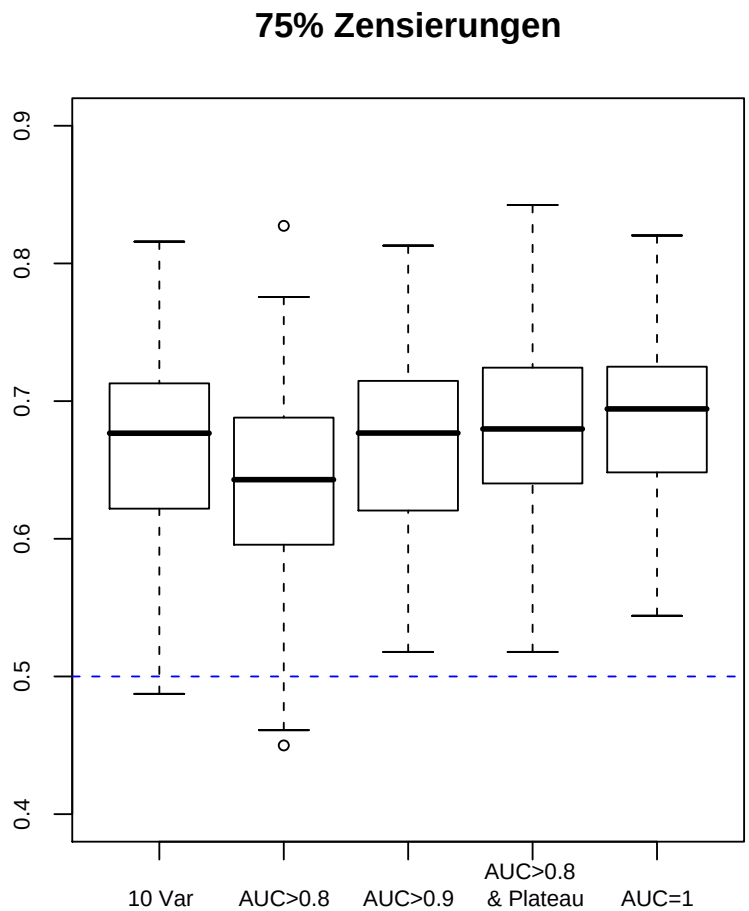




Unabhängige Einflussgrößen

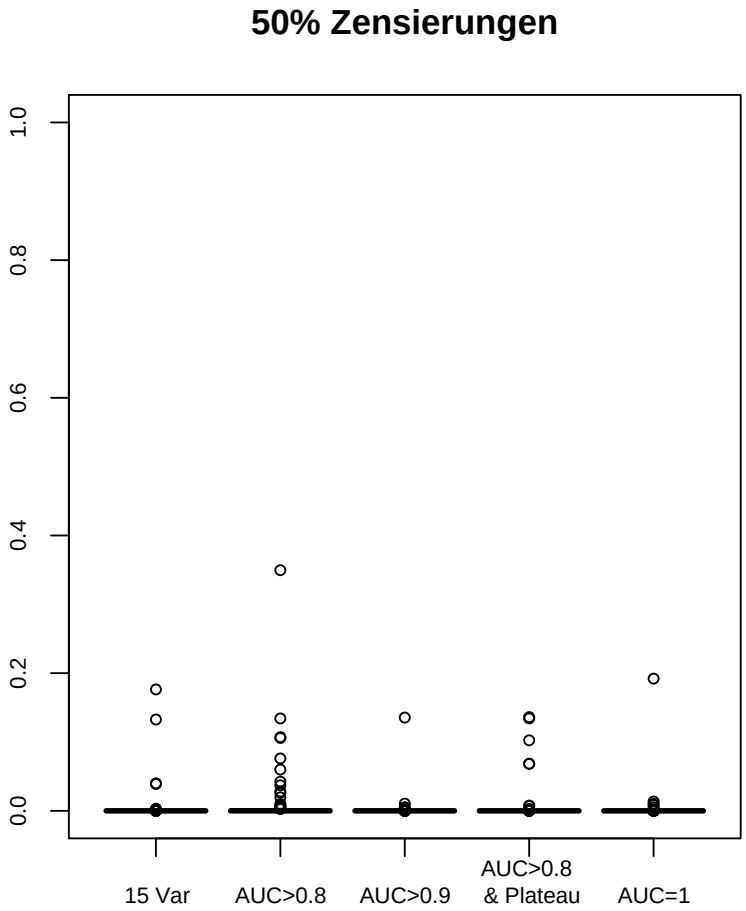
AUCs des Scores auf den Testdaten

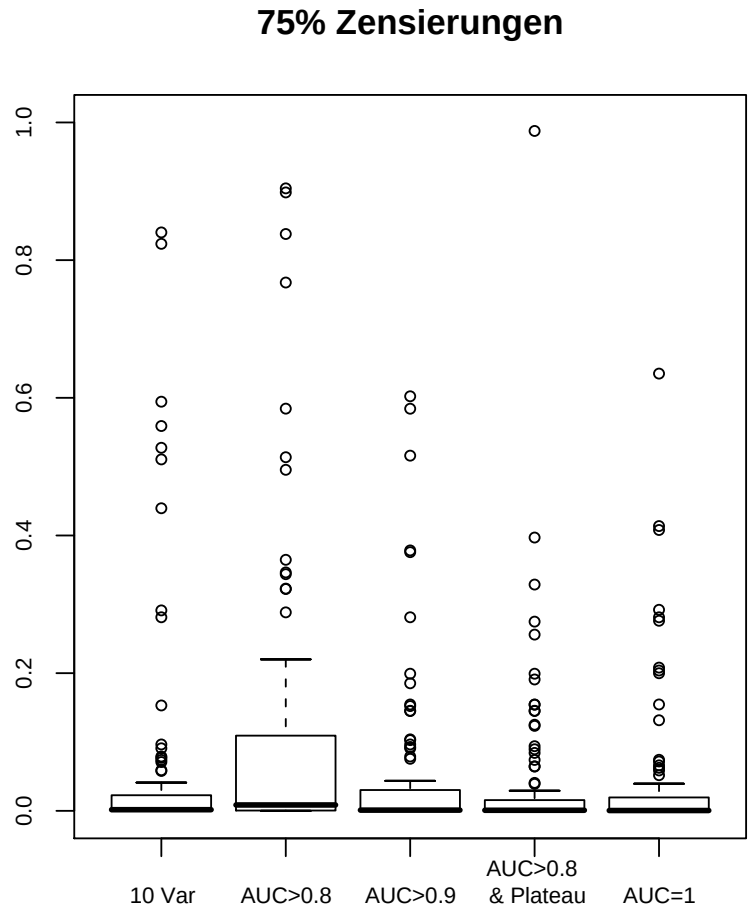




Unabhängige Einflussgrößen

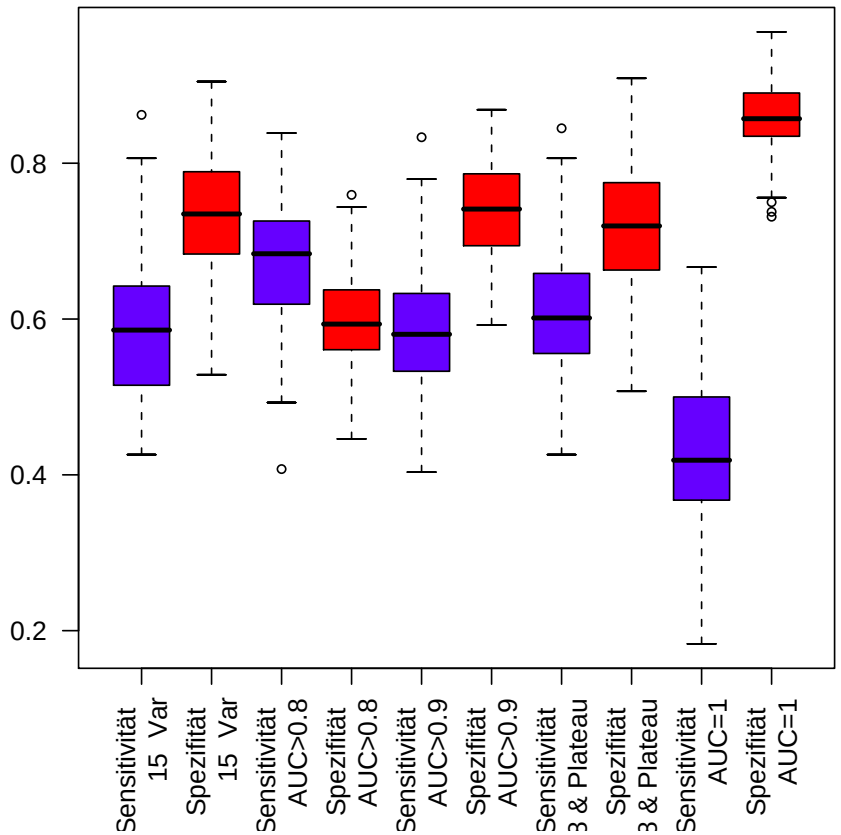
p-Werte für gute vs. schlechte Prognose, 80% Sensitivität, Testdaten





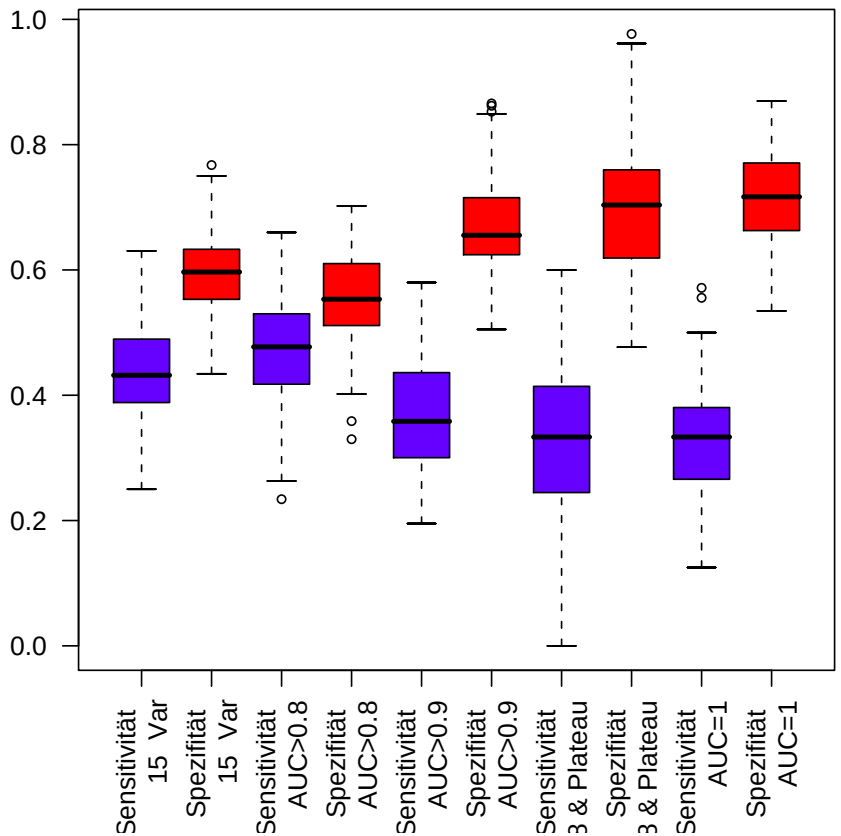
Unabhängige Einflussgrößen

50% Zensierungen, 80%Sensitivität, Testdaten



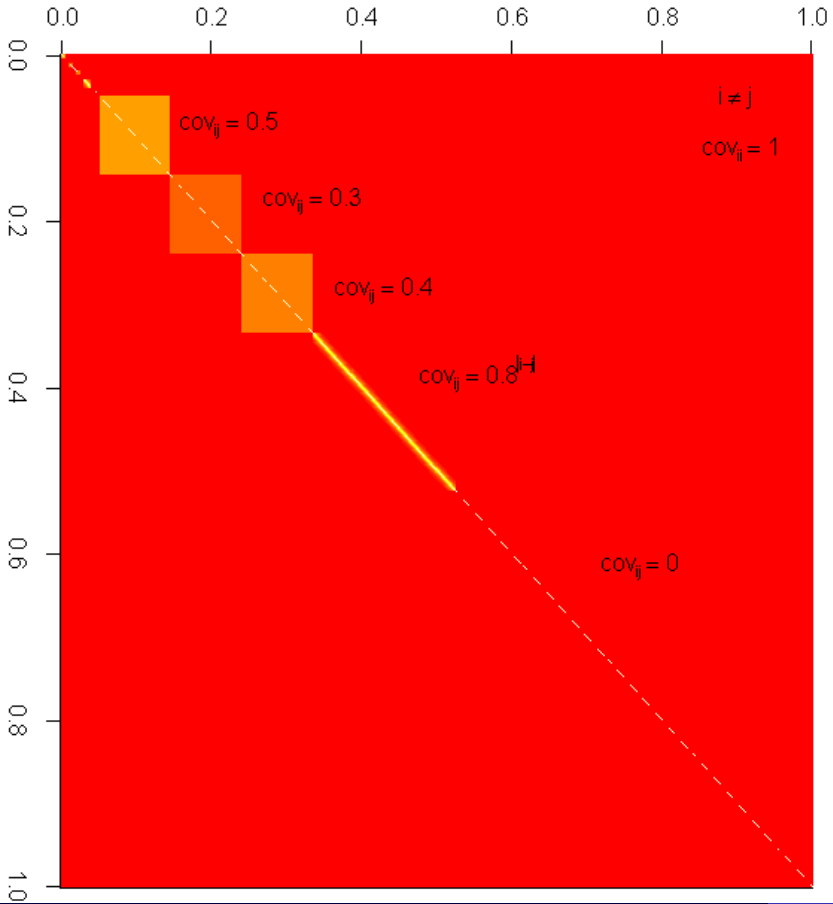
Unabhängige Einflussgrößen

75% Zensierungen, 80%Sensitivität, Testdaten



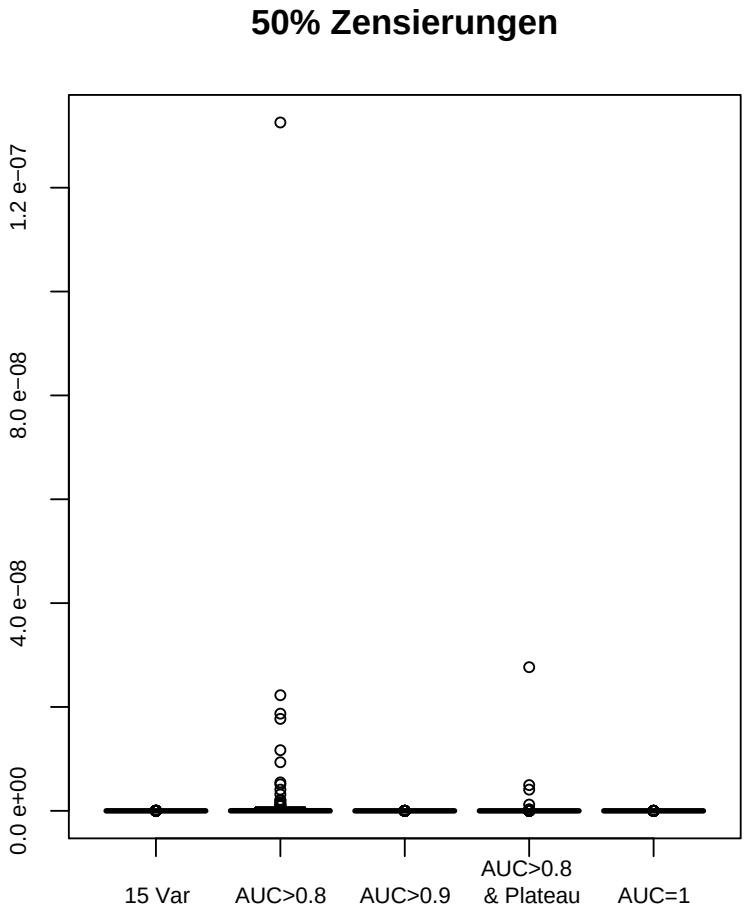
Abhängige Einflussgrößen

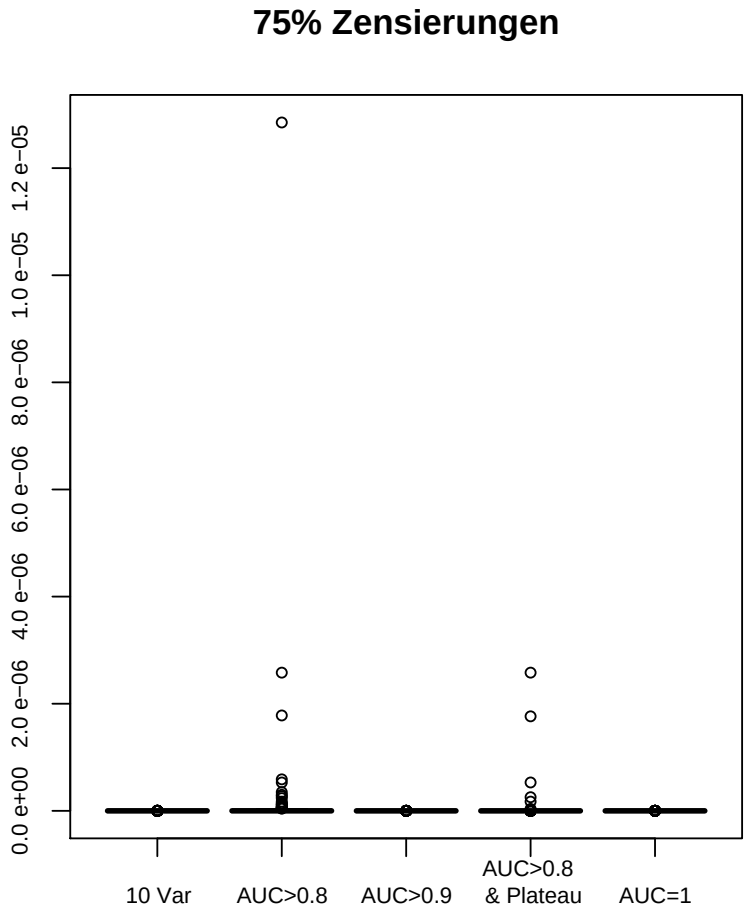
Kovarianzmatrix



Abhängige Einflussgrößen

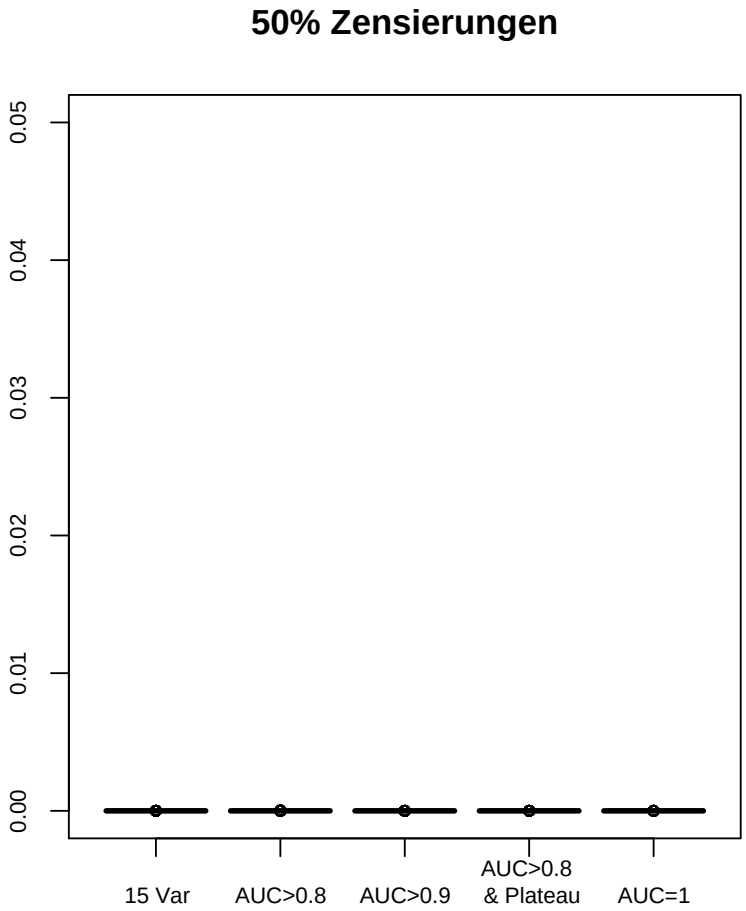
p-Werte für gute vs. schlechte Prognose, 80% Sensitivität,Lerndaten

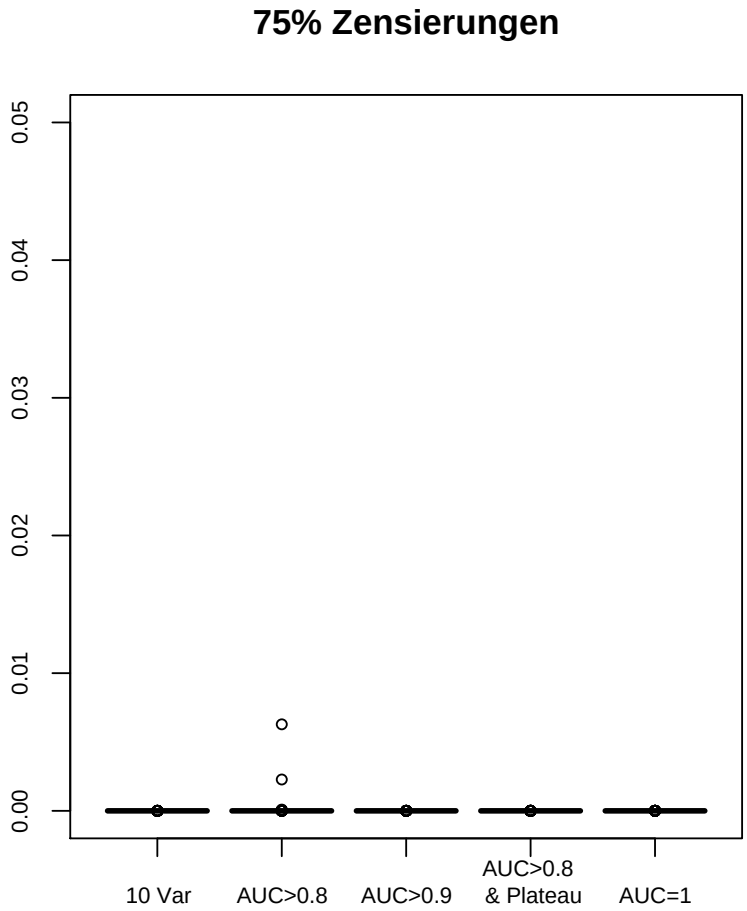




Abhängige Einflussgrößen, Test

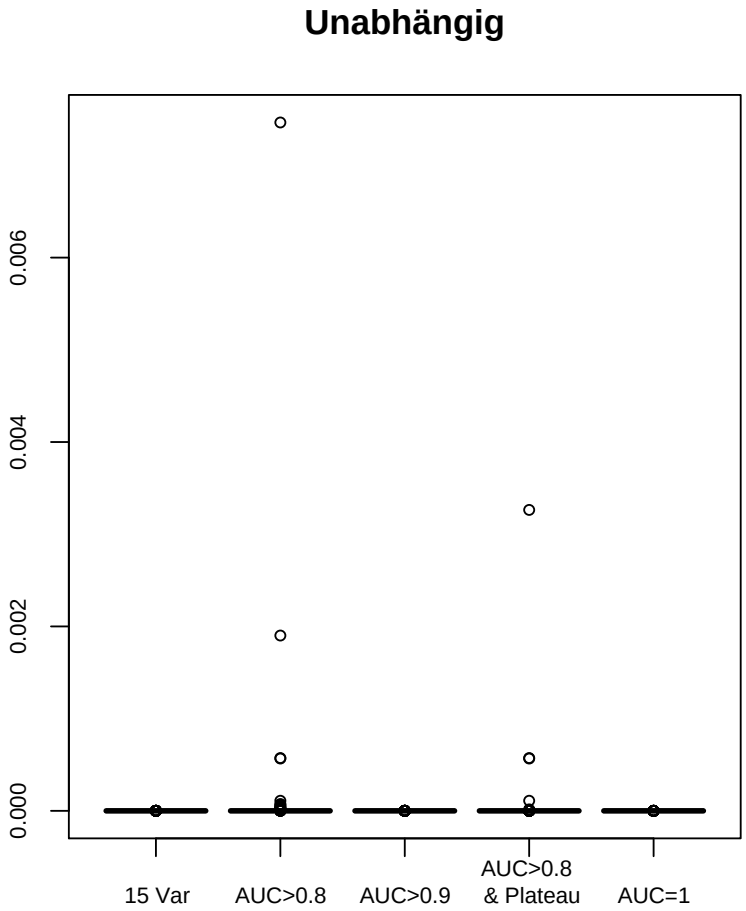
p-Werte für gute vs. schlechte Prognose, 80% Sensitivität, Testdaten

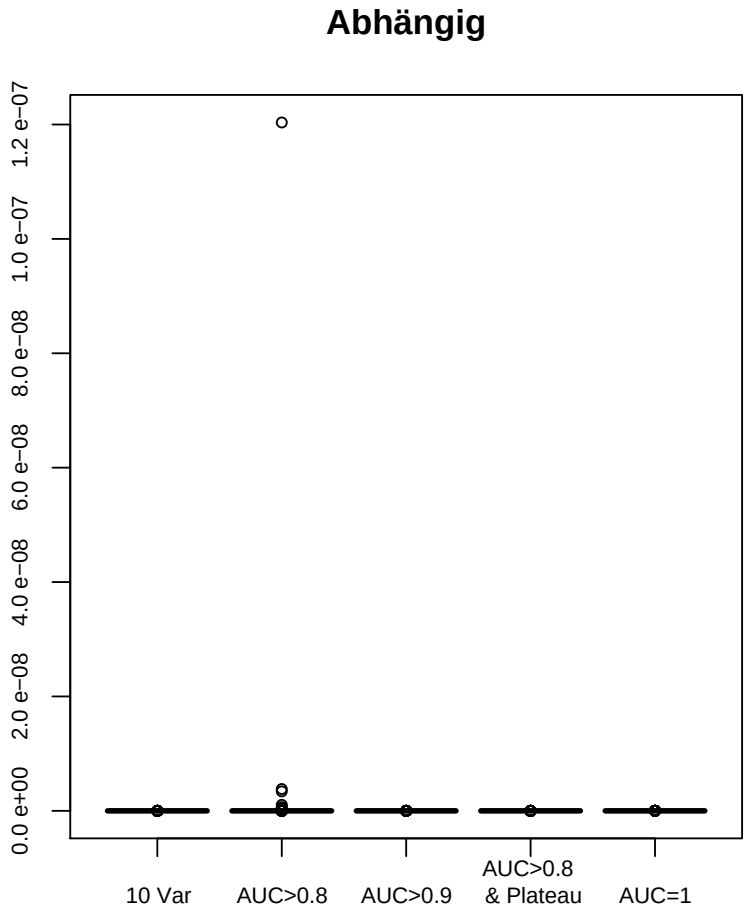




Schwächerer Zusammenhang

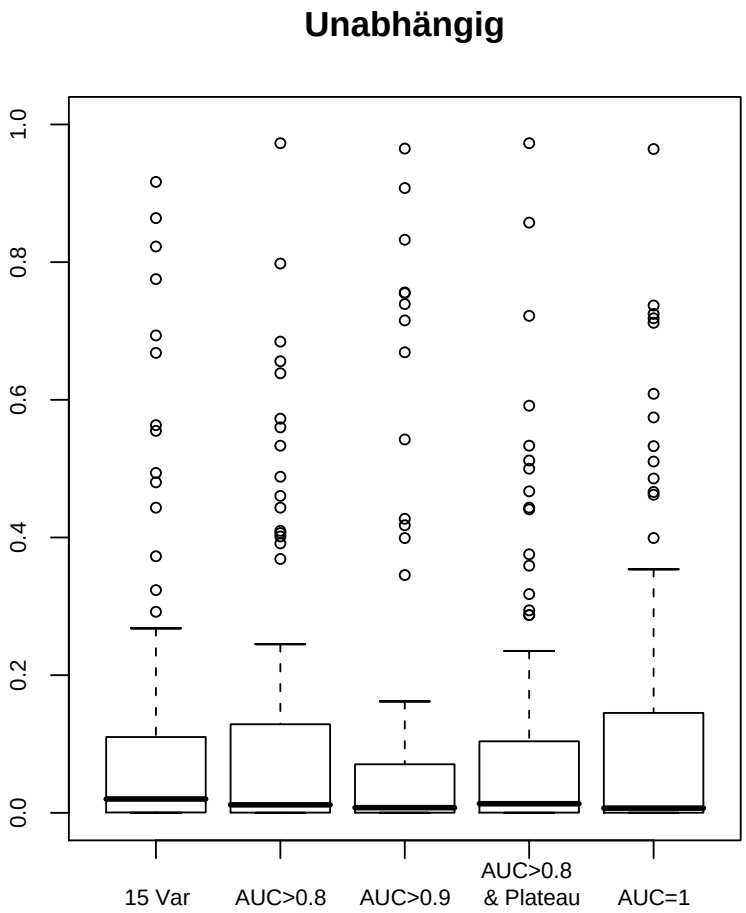
p-Werte für gute vs. schlechte Prognose, 80% Sensitivität, Lerndaten

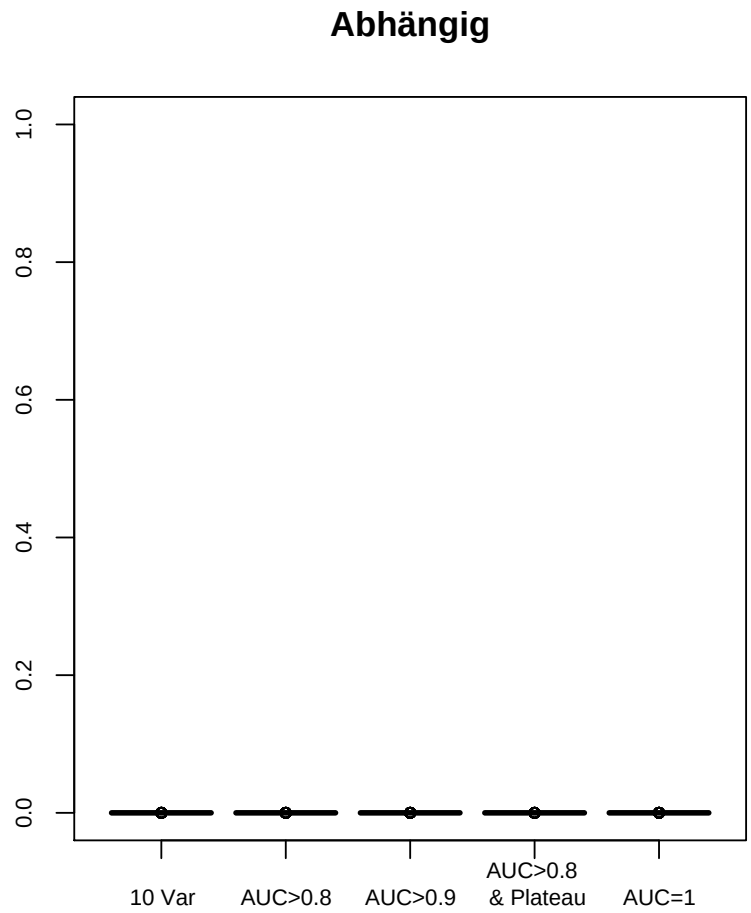




Schwächerer Zusammenhang, Test

p-Werte für gute vs. schlechte Prognose, 80% Sensitivität, Testdaten





Zusammenfassung der Ergebnisse

- Validierung des Scores von Wang et al. relativ erfolgreich (90% Sensitivität, 50% Spezifität) (Foekens et al, J Clin Oncol 2006)
- bei KHK-Daten eher mäßige Ergebnisse
- hoher Anteil Zensierungen führt zu schlechten Ergebnissen
- zu hohe Sensitivität bei Wahl der Schwelle nicht sinnvoll
- positive Korrelation der Einflussgrößen hier kein Problem
- bei höherem Informationsgehalt der Daten bessere Ergebnisse

Diskussion

- Dichotomisierung der Survival-Information problematisch (Wahl des Zeitpunktes)
- Wie groß soll die AUC sein / Wieviele Variablen sind nötig?
- Wahl der Anzahl Variablen ist datengesteuert
- Wo wählt man die Schwelle zur Einteilung gute / schlechte Prognose?
- Risiko des Overfit
- Abhängige Einflussgrößen?

Literatur

- Yixin Wang, Jan G.M. Klijn et al.
  Gene expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer
  Lancet, 2005
- John A. Foekens, David Atkins et al.
  Multicenter Validation of a Gene Expression-Based Prognostic Signature inLymph Node-Negative Primary Breast Cancer
  J Clin Oncol, 2006
- Ralf Bender, Thomas Augustin, Maria Blettner
  Generating survival times to simulate Cox proportional hazards models
  Statistics in Medicine, 2005