

## 7 Anhang

### 7.1 Abkürzungen und Spezialbegriffe

<b>CART</b>	Klassifikations- und Regressions-Baum ( <b>C</b> lassification and <b>R</b> egression <b>T</b> ree) [19]
<b>Cox-PH-Regression</b>	Proportionale Hazard-Regression nach Cox [38]
<b>Design</b>	S-Plus-Bibliothek mit speziellen Funktionen zur Entwicklung und Validierung von multivariaten Modellen (Frank Harrell)
<b>Hmisc</b>	S-Plus-Bibliothek mit vielen nützlichen Funktionen (Frank Harrell)
<b>KNN</b>	Abkürzung für <b>K</b> ünstliche <b>N</b> euronale <b>N</b> etze – [175]
<b>ROC</b>	<b>R</b> eceiver <b>O</b> peration <b>C</b> haracteristic
<b>Rpart</b>	Rekursive Partitionierung. Der Name für die S-Plus-Bibliothek von T. Therneau wurde gewählt, weil der Name CART schon benutzt wurde. Bei Rpart handelt es sich aber um erweiterte Funktionen zur Generierung von CART
<b>SurvNN</b>	Diese S-Plus-Bibliothek wurde im Rahmen der Dissertation von Ruth M. Ripley entwickelt und an mehreren Beispielen getestet [134, 135]. Sie basiert in wesentlichen Punkten auf der Bibliothek NNet von Brian Ripley.
<b>Impute</b>	Der in der englischsprachigen Literatur häufig zu findende Begriff <i>Imputation</i> wurde in dieser Arbeit mit Zuschreibung übersetzt. Gemeint ist die Ersetzung fehlender Angaben durch Werte, die mit statistischen Verfahren geschätzt wurden. Genauere Angaben dazu finden sich im Methodenteil.

## 7.2 **Abbildungsverzeichnis**

Abbildung 1: Vorwärtsgerichtetes zweistufiges neuronales Netz .....	9
Abbildung 2: Neuron.....	9
Abbildung 3: Beispiel für ein neuronales Netz zur Präsentation der XOR-Funktion .....	11
Abbildung 4: Netze ohne Rückkopplung .....	12
Abbildung 5: Beispiel einer Kaplan-Meier-Überlebenskurve .....	18
Abbildung 6: Modell der Haut mit den verschiedenen Hautschichten; Darstellung des Invasionslevels nach Clark und der Tumordicke nach Breslow (modifiziert nach Fritsch [67]) .....	28
Abbildung 7: Überlebenskurven für die Variable F1 der simulierten Daten (Trainigsdaten).....	39
Abbildung 8: Überlebenskurven für die Variable F1 der simulierten Daten (Testdaten).....	39
Abbildung 9: Überlebenskurven für die Variable F2 der simulierten Daten (Trainigsdaten).....	40
Abbildung 10: Überlebenskurven für die Variable F2 der simulierten Daten (Testdaten).....	40
Abbildung 11: Überlebenskurven für die Variable F3 der simulierten Daten (Trainigsdaten).....	41
Abbildung 12: Überlebenskurven für die Variable F3 der simulierten Daten (Testdaten).....	41
Abbildung 13: Überlebenskurven für die Variable F4 der simulierten Daten (Trainigsdaten).....	42
Abbildung 14: Überlebenskurven für die Variable F4 der simulierten Daten (Testdaten).....	42
Abbildung 15: Überlebenskurven für die Variable F5 der simulierten Daten (Trainigsdaten).....	43
Abbildung 16: Überlebenskurven für die Variable F5 der simulierten Daten (Testdaten).....	43
Abbildung 17: Komplexität des CART-Baumes und relativer Fehler bestimmt durch Kreuzvalidierungen.....	44
Abbildung 18: CART-Baum für die simulierten Daten unter Berücksichtigung von zensierten Beobachtungen.....	45
Abbildung 19: Überprüfung der Proportionalitätsannahme für die Variable F2.....	46
Abbildung 20: Überprüfung der Proportionalitätsannahme für die Variable F3.....	47
Abbildung 21: Log-Relative Hazards für die Variablen F1-F4, adjustiert für: F1=1.253, F2=2; F3=6.462, F4=1.468 .....	47
Abbildung 22: Struktur des verwendeten Netzwerks für die Analyse der simulierten Daten .....	48
Abbildung 23: ROC-Kurven für Prognoseschätzung bei den simulierten Daten mit einem Parameter (Variable F3) .....	51
Abbildung 24: ROC-Kurven für die Prognoseschätzung bei den simulierten Daten mit einem CART-Baum.....	52
Abbildung 25: ROC-Kurven für die Prognoseschätzung bei den simulierten Daten mit einem Cox-PH-Modell .....	53
Abbildung 26: ROC-Kurven für die Prognoseschätzung bei den simulierten Daten mit einem künstlichen neuronalen Netz .....	54
Abbildung 27: Überlebenskurven für die Variable Tumordicke .....	63
Abbildung 28: Überlebenskurven für die Variable Ulzeration .....	64
Abbildung 29: Überlebenskurven für die Variable Geschlecht .....	64
Abbildung 30: Überlebenskurven für die Variable Invasionslevel .....	65

---

Abbildung 31: Überlebenskurven für die Variable Alter .....	65
Abbildung 32: Überlebenskurven für die Variable Lokalisation .....	66
Abbildung 33: Überlebenskurven für die Variable histologischer Typ.....	66
Abbildung 34: Darstellung der optimalen Baumgröße basierend auf einem Parameter für die Komplexität.....	67
Abbildung 35: CART-Baum mit sieben Knoten für die Melanomdaten (TD steht für die Tumordicke nach Breslow) .....	68
Abbildung 36: Schoenfeld-Residuen für die Tumordicke.....	70
Abbildung 37: Log Relative Hazards für die untersuchten Faktoren.....	71
Abbildung 38: Kalibrierung des Cox-PH-Modells mit jeweils 1000 Patientendaten für jede Gruppe.....	73
Abbildung 39: ROC-Kurven für Prognoseschätzung bei Registerdaten mit einem Parameter (Tumordicke) .....	76
Abbildung 40: ROC-Kurven für Prognoseschätzung bei Registerdaten mit einem CART-Baum.....	77
Abbildung 41: ROC-Kurven für Prognoseschätzung bei Registerdaten mit einem Cox-PH-Modell.....	78
Abbildung 42: ROC-Kurven für die Prognoseschätzung bei Registerdaten mit einem künstlichen neuronalen Netz .....	79
Abbildung 43: ROC-Kurven für die Prognoseschätzung bei Registerdaten mit einem komplexeren künstlichen neuronalen Netz.....	80
Abbildung 44: Darstellung von Fehlerrate und Komplexität des CART-Baums für die vervollständigten Registerdaten.....	86
Abbildung 45: CART-Baum für die Registerdaten mit vervollständigten Angaben.....	87
Abbildung 46: ROC-Kurven für Prognoseschätzung bei vervollständigten Registerdaten mit der Tumordicke .....	90
Abbildung 47: ROC-Kurven für Prognoseschätzung bei vervollständigten Registerdaten mit einem CART-Baum .....	91
Abbildung 48: ROC-Kurven für Prognoseschätzung bei vervollständigten Registerdaten mit einem Cox-Modell.....	92
Abbildung 49: ROC-Kurven für Prognoseschätzung bei vervollständigten Registerdaten mit einem künstlichen neuronalen Netz.....	93
Abbildung 50: ROC-Kurven für Prognoseschätzung bei vervollständigten Registerdaten mit einem komplexen künstlichen neuronalen Netz .....	94

### 7.3 Tabellenverzeichnis

Tabelle 1 : Ausgaben der XOR-Funktion .....	10
Tabelle 2: Fiktive Verlaufsdaten von 10 Personen .....	17
Tabelle 3: Kategorien für die TNM-Klassifikation.....	29
Tabelle 4: Stadieneinteilung des malignen Melanoms und 5-Jahres-Überlebenswahrscheinlichkeit.....	30
Tabelle 5: Arbeitsschritte bei der Analyse der Daten .....	31
Tabelle 6: Gewichte des neuronalen Netzes .....	49
Tabelle 7: Rang-Korrelationen für die Trainingsdaten und die verschiedenen Modelle bei den simulierten Daten (4000 Fälle, davon 474 nicht zensiert und 2251972 relevante Paare).....	55
Tabelle 8: Rang-Korrelationen für die Trainingsdaten und die verschiedenen Modelle bei den simulierten Daten (2000 Fälle, davon 240 nicht zensiert und 562994 relevante Paare).....	55
Tabelle 9: Verteilung der Patienten aus den ausgewählten acht Kliniken auf die drei Analysegruppen.....	57
Tabelle 10: Verteilung der untersuchten Faktoren bei 8908 Patienten .....	58
Tabelle 11: Verteilung der untersuchten Faktoren bei 7780 Patienten (Patienten mit fehlenden Angaben ausgeschlossen).....	60
Tabelle 12: Untersuchte Variablen, Fallzahlen und Ereignisse .....	62
Tabelle 13: Wald-Statistik für das Cox-PH-Modell .....	69
Tabelle 14: Ergebnisse der Validierung des COX-PH-Modells .....	72
Tabelle 15: Ergebnisse der Kalibrierung des COX-PH-Modells .....	73
Tabelle 16: Netzgewichte des einfachen künstlichen neuronalen Netzes .....	74
Tabelle 17: Netzgewichte des komplexen künstlichen neuronalen Netzes .....	74
Tabelle 18: Rang-Korrelationen für die Trainingsdaten und die verschiedenen Modelle bei 4021 Patienten, davon 306 nicht zensiert und 1460610 relevante Paare.....	81
Tabelle 19: Rang-Korrelationen für die Testdaten und die verschiedenen Modelle bei 2022 Patienten, davon 161 nicht zensiert und 361364 relevante Paare.....	81
Tabelle 20: Fehlende Angaben bei den Faktoren .....	83
Tabelle 21: Verteilungen der Variablen in den Trainingsdaten. Bei den vervollständigten Werten fehlen die Fälle ohne Angabe zur Tumordicke.....	84
Tabelle 22: Wald-Statistik für das Cox-PH-Modell bei der Analyse von Registerdaten mit vervollständigten Angaben.....	88
Tabelle 23: Rang-Korrelationen für die Trainingsdaten und die verschiedenen Modelle bei den vervollständigten Registerdaten (5460 Fälle, davon 506 nicht zensiert, 3416962 relevante Paare).....	95
Tabelle 24: Rang-Korrelationen für die Trainingsdaten und die verschiedenen Modelle bei den vervollständigten Registerdaten (2743 Fälle, davon 259 nicht zensiert, 828370 relevante Paare).....	95

## 7.4 Programm zur Datensimulation

### \* Generierung von Zufallszahlen.

```
compute status=0.
compute id=$casenum.
compute v1=rv.normal(1,0.5).
compute v2=rv.uniform(1,5).
compute v3=rv.uniform(1,10).
compute Zeit = rv.weibull(1,1)*12.
compute zz1 = rv.uniform(0,1).
compute zz2 = rv.uniform(0,1).
compute zz3 = rv.uniform(0,1).
```

### \* Erzeugung der Variablen Status und Nachbeobachtungszeit.

```
if (v1+v2 > 5) and (zz1 <= 0.3) Status=1.
if (v1+v2 > 5) and (zz1 <= 0.5) Zeit=Zeit*(2/3).
if (v1+v3 > 10) and (zz2 <= 0.4) Status=1.
if (v1+v3 > 10) and (zz2 <= 0.4) Zeit=Zeit*(2/3).
if (v2*v3 > 40) and (zz3 <= 0.5) Status=1.
if (v2*v3 > 40) and (zz3 <= 0.5) Zeit=Zeit*(2/3).
```

### \* Erzeugung von Variablen/Hinzufügen von Rauschen.

```
compute f1=v1+v4/2.
compute f2=trunc(v2).
compute f3=v3+v5.
compute f4=v6.
compute f5=trunc(v1).
execute.
```

## 7.5 Beispiele für S-Plus-Auswertungsprogramme

### # Berechnung eines CART

```
library(RPart)
fit <- rpart(Surv(NACHZT, STATUS)
            ~TD+LEVEL+TANS+HISTONEU+GESCHL+ALTER+HULZ,
            data=training.om, control=rpart.control(cp=0.0001))
plotcp(fit)
fit1->prune(fit, cp=0.004)
plot(fit1)
text(fit1)
post(fit1, file="")

training.om$pred.rpart<-predict(fit1)
test.om$pred.rpart<-predict(fit1, newdata=test.om)
```

```
# COX-PH-Analyse
library(hmisc,T)
library(Design,T)
attach(training.om)
GESCHLN<-factor(GESCHL)
HISTON<-factor(HISTONEU)
HULZN<-factor(HULZ)
TANSN<-factor(TANS)
LEVELN<-ordered(LEVEL)
TDN<-log(TD)
dd<-datadist(GESCHLN,TDN,ALTER,HULZN,HISTON,TANSN,LEVELN)
options(datadist='dd')
S<-Surv(NACHZT,STATUS)
f<-cph(S~GESCHLN+rcs(TDN,4)+HULZN+HISTON+LEVELN+TANSN+ALTER,
      eps=0.000001,x=T,y=T,surv=T,time.inc=60)
anova(f)
fastbw(f)
z<-predict(f,type='terms')
fs<-cph(S~z,x=T,y=T)
phtest<-cox.zph(fs,transform='identity')
plot(phtest,var='TDN')

par(mfrow=c(2,2))
plot(f,GESCHLN=NA)
plot(f,TDN=NA)
plot(f,HULZN=NA)
plot(f,HISTON=NA)

par(mfrow=c(2,2))
plot(f,LEVELN=NA)
plot(f,TANSN=NA)
plot(f,ALTER=NA)

validate(f,B=200,dxy=T)
calil1<-calibrate(f,B=500,u=60,m=1000)

training.om$pred.cox<-predict(f,)
detach()
```

```
# KNN
library(hmisc,T)
library(Design,T)
library(survnnnet,T)
attach(training.om)
GESCHLN<-factor(GESCHL)
HISTON<-factor(HISTONEU)
HULZN<-factor(HULZ)
TANSN<-factor(TANS)
LEVELN<-ordered(LEVEL)
TDN<-log(TD)
dd<-datadist(GESCHLN,TDN,ALTER,HULZN,HISTON,TANSN,LEVELN)
options(datadist='dd')
S<-Surv(NACHZT,STATUS)
nn<-phnnet(S~GESCHLN+TDN+HULZN+HISTON+LEVELN+TANSN+ALTER,
           decay = 0.01,size=3,skip=T,trace=T,maxit=1000,bias.decay=25)
training.om$pred.nn2 <-predict(nn)

detach()
attach(test.om)
GESCHLN<-factor(GESCHL)
HISTON<-factor(HISTONEU)
HULZN<-factor(HULZ)
TANSN<-factor(TANS)
LEVELN<-ordered(LEVEL)
TDN<-log(TD)
dd<-datadist(GESCHLN,TDN,ALTER,HULZN,HISTON,TANSN,LEVELN)
options(datadist='dd')
S<-Surv(NACHZT,STATUS)
test.om$pred.nn2 <-predict(nn,newdata=test.om)
```



**# Rangkorrelationen für zensierte Daten**

```
attach(training.mm)
S<-Surv(NACHZT, STATUS)
rcorr.cens(-pred.td, S)
rcorr.cens(-pred.rpart, S)
rcorr.cens(-pred.cox, S)
rcorr.cens(-pred.nn, S)
rcorr.cens(-pred.nn2, S)
detach()
```

```
attach(test.mm)
S<-Surv(NACHZT, STATUS)
rcorr.cens(-pred.td, S)
rcorr.cens(-pred.rpart, S)
rcorr.cens(-pred.cox, S)
rcorr.cens(-pred.nn, S)
rcorr.cens(-pred.nn2, S)
detach()
```