

Ein Konfidenzband für ROC-Kurven mit SAS

Steffen Witte
Universität Heidelberg
Abt. Med. Biometrie
Im Neuenheimer Feld 305
69120 Heidelberg
witte@imbi.uni-
heidelberg.de

Katrin Jensen
Universität Heidelberg
Abt. Med. Biometrie
Im Neuenheimer Feld 305
69120 Heidelberg
jensen@imbi.uni-
heidelberg.de

Zusammenfassung

Bei einem quantitativen diagnostischen Test hängen Spezifität (Anteil der korrekt klassifizierten Gesunden) und Sensitivität (Anteil der korrekt klassifizierten Kranken) von der Wahl des Schwellenwertes ab. Durchläuft der Schwellenwert den gesamten Bereich möglicher Testwerte, und wird jeweils Sensitivität gegen Spezifität aufgetragen, so entsteht die sogenannte ROC-Kurve (receiver operating characteristic curve). Die ROC-Kurve ist eine Visualisierung der diagnostischen Trennschärfe des Tests über den gesamten Testwert-Bereich.

In der Praxis muss die ROC-Kurve eines quantitativen diagnostischen Tests anhand empirischer Daten geschätzt werden. Sowohl die Berechnung der Spezifitäten und Sensitivitäten als auch die Darstellung der empirischen ROC-Kurve selbst sind mit SAS problemlos möglich. Um diese empirische ROC-Kurve jedoch zur Unterstützung bei der Wahl eines geeigneten Schwellenwertes heranziehen zu können, muss die Zufallsvariabilität geeignet dargestellt werden. Zahlreiche Vorschläge für punktweise und simultane Konfidenzbänder für ROC-Kurven sind in der Literatur beschrieben und diskutiert; allerdings sind die vorgeschlagenen Konfidenzbänder weder in SAS noch mit anderen gängigen Statistikprogrammen zu erstellen.

In diesem Beitrag soll ein Ansatz für punktweise Konfidenzbänder dargestellt und illustriert werden, der von Hilgers vorgeschlagen wurde. Ein SAS Macro zur Darstellung der Konfidenzbänder von Hilgers wurde entwickelt.

Abstract

Considering a quantitative diagnostic test procedure specificity (rate of correct classified healthy people) and sensitivity (rate of correct classified sick people) depend on the cut-off. If the cut-off will be shifted through the range of possible test values and the resulting sensitivities and specificities are plotted, the graph results in a ROC-curve (receiver operating characteristic curve). Therefore the ROC-curve is a visualisation of the power of the diagnostic test to distinguish between a diseased and non-diseased individual on the whole range of test values. For practical purposes the ROC-curve must be estimated and plotted which is easy with SAS. Anyway, to use the

ROC-curve to support the choice of the best cut-off, the variability must be incorporated. Several approaches are available for pointwise and simultaneous confidence bounds but none of them are implemented in SAS or other usual statistical software packages. In this paper the pointwise confidence bounds introduced by Hilgers will be discussed. A SAS macro for the graphical output will illustrate the results.

Keywords: ROC, SAS, macro, Konfidenzband, diagnostics.

1 Einleitung

Receiver operating characteristics, oder kurz ROC, stammen aus dem Bereich der Entscheidungstheorie und sind in der Biometrie deutlich weiter entwickelt worden. Eine einfache Literatursuche in MEDLINE, einer in der Biometrie üblichen Datenbank, liefert deutliche Zuwachsraten von Publikationen mit oder über ROC-Kurven im Verhältnis zu der Anzahl gesamtter Einträge in der Datenbank. Die relativen Häufigkeiten für die Jahre 1990-2002 sind in Abbildung 1 dargestellt.

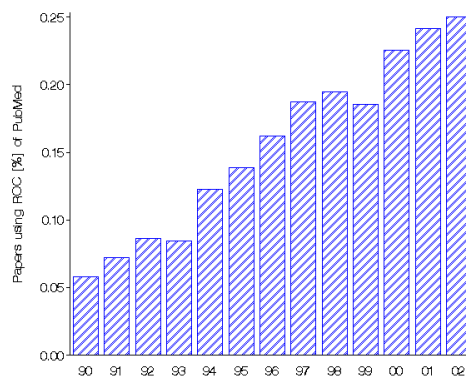


Abbildung 1: Relative Häufigkeit von Publikationen die ROC verwenden.

Motivation:

Die Methodik von ROC-Kurven kann in verschiedenen Bereichen auftreten. Folgende Beispiele sollen dies illustrieren. Biometrie-Beispiel: Der Krankheitsstatus eines Menschen (gesund/krank) soll mit Hilfe eines Testverfahrens diagnostiziert werden. Das Testverfahren liefert einen ordinalen oder sogar stetigen Messwert. Kredit-Beispiel: Das Ausfallrisiko eines Kredites (Rückzahlung: ja/nein) soll anhand einer Kreditwürdigkeitsprüfung mithilfe eines multiplen Scores (ordinal) beurteilt werden. Material-Beispiel: Zur Untersuchung eines Scanverfahrens (stetig), dass zur Überprüfung von Materialschäden eingesetzt werden soll, wird zur Validierung des Verfahrens einigen Materialstücken ein Fehler hinzugefügt (Fehler liegt vor/liegt nicht vor). In allen drei Fällen geht es um die Bewertung eines Testverfahrens mit ordinalen oder stetigen Messgrößen zur Ermittlung eines binären Zustandes (Zielgröße).

Cut-off, Sensitivität, Spezifität:

Das Testverfahren liefert einen Messwert, der kategorisiert werden muss, um den binären Zustand zu prognostizieren. Anhand eines cut-offs c wird entschieden, ob ein positives oder negatives Testergebnis vorliegt. Hier soll ein Messwert $> c$ ein positives Testergebnis bedeuten (T+) und ein Messwert $\leq c$ ein negatives Testergebnis (T-). Der binäre Zustand wird, angelehnt an das biometrische Beispiel, mit K+ (=krank) oder K- (=gesund) bezeichnet. Zur Bewertung eines Testverfahrens an einem bestimmten cut-off dienen Sensitivität und Spezifität. Die Sensitivität ist die Wahrscheinlichkeit einen Kranken mit dem Test auch als krank zu erkennen = $P(T+ | K+)$. Die Spezifität ist die Wahrscheinlichkeit einen Gesunden mit dem Test auch als gesund zu erkennen = $P(T- | K-)$. Die beiden bedingten Wahrscheinlichkeiten können aus einem Datensatz einfach geschätzt werden. Das folgende fiktive Beispiel in Tabelle 1 mit 60 Personen, 25 Kranken und 35 Gesunden, soll die Schätzung der Sensitivität und der Spezifität verdeutlichen. Hier sind die Schätzer $Sens=19/25$ und $Spec=30/35$.

Tabelle 1: Beispiel für einen cut-off: Sensitivitäts- und Spezifitätsschätzung

	K+	K-	Summe
T+	19	5	24
T-	6	30	36
Summe	25	35	60

Die Messdaten des zu validierenden Tests sind Realisierungen von Zufallsvariablen mit einer unbekanntem Verteilung. Um die Sensitivität und die Spezifität graphisch zu verdeutlichen wurden in Abbildung 2 die Dichten der Gesunden (K-) und Kranken (K+) dargestellt und ein cut-off gewählt. Die schraffierten Flächen stellen dann die Sensitivität bzw. die Spezifität des diagnostischen Tests mit dem entsprechenden cut-off dar.

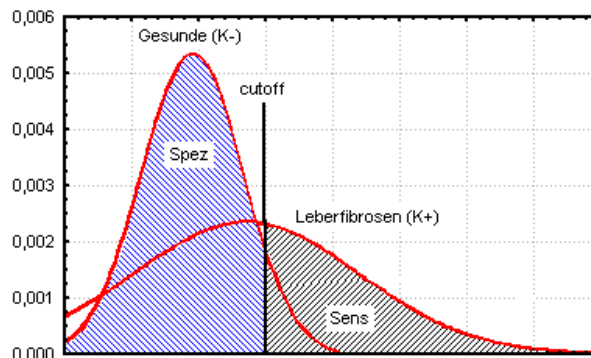


Abbildung 2: Sensitivität und Spezifität als Flächen unter Dichten

ROC-Kurve:

Die graphische Darstellung von Spezifität (x-Achse) und Sensitivität (y-Achse) für alle möglichen cut-offs bildet eine ROC-Kurve. Zur Schätzung sollen hier die nichtparametrischen Schätzer der beiden Maße für alle Datenpunkte (alle möglichen cut-offs) verwendet werden. Wie auf der 6. KSFE vorgestellt wurde, können zum Beispiel mit Hilfe des Macros %rocn alle notwendigen Schätzer ermittelt und anschließend mit %rocplot graphisch dargestellt werden. Das Daten-Beispiel aus der Arbeit von Schäfer soll hier verwendet werden (Schäfer 1994).

```
%rocn(inset=1nl dh, event=diag, tests=1nl dh, outset2=1nl dh2);  
filename ausgabe '1dh-line1.pdf';  
%rocplot(gfile=ausgabe, gdev=pdf, inset=1nl dh2, by=_test);
```

liefert die Abbildung 3 als pdf-Datei.

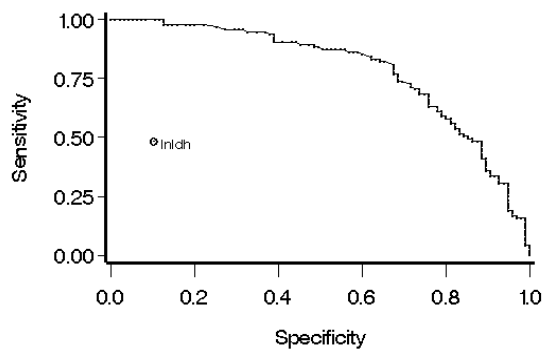


Abbildung 3: ROC-Kurve

Zielsetzung:

Die Darstellung der ROC-Kurve wie in Abbildung 3 entspricht einem Punktschätzer und sollte durch eine Konfidenzaussage gestützt werden. Insbesondere für die spätere Interpretation werden die unteren Grenzen von Konfidenzbändern verwendet, um eine Mindestsensitivität bei vorgegebener Spezifität zu gewährleisten. Daher ist es notwendig, ein Konfidenzband um die ROC-Kurve zu legen. Verwendet werden soll dazu das Verfahren von Hilgers (Hilgers 1991).

2 Konfidenz

Um eine Konfidenzaussage zu treffen, sollen folgende drei Typen unterschieden werden.

- Punktweise Konfidenzintervalle für die Sensitivität an einem bestimmten cut-off c basieren auf der Binomialverteilung und können mit dem Macro %cibinomexact berechnet werden (Daly 1992). Hier geht dann nur der Datensatz mit den Kranken (K+) ein:
 $P(LCL(c) \leq Sens(c) \leq UCL(c)) = 95\%$ für alle cut-offs
- Punktweise Konfidenzbänder halten das Niveau für eine bestimmte tatsächliche Spezifität ein. Dazu gibt es vor allem die Methoden nach Hilgers (Hilgers 1991) und Schäfer (Schäfer 1994):
 $P(LCL(Spec) \leq Sens(Spec) \leq UCL(Spec)) = 95\%$ für alle Spec
- Simultane Konfidenzbänder halten das Niveau simultan für alle Spezifitäten ein und liefern einen Vertrauensbereich simultan für alle Sensitivitäten. Die Methoden nach Campbell (Campbell 1994) und Jensen (Jensen 2000) liefern simultane Konfidenzbänder mit:
 $P(LCL(Spec) \leq Sens(Spec) \leq UCL(Spec)) = 95\%$ für alle Spec

Die Interpretationsmöglichkeiten der drei verschiedenen Typen ergeben sich direkt aus den Definitionen und erst die simultanen Konfidenzbänder verdienen den Namen ein Konfidenzband für die (gesamte) ROC-Kurve zu liefern. Aber für die häufige biometrische Fragestellung eine Mindestsensitivität bei vorgegebener Spezifität zu gewährleisten, reicht ein punktweises Konfidenzband aus.

3 Umsetzung in SAS: %roccih und %rocplot

Methodik:

In dem Artikel von Schäfer wird die Methode nach Hilgers auch als "two-stage confidence bounds" beschrieben (Schäfer 1994). Betrachtet man einen beliebigen cut-off und seien die Sensitivität und Spezifität bereits geschätzt, so verwendet man im ersten Schritt die Daten der Gesunden (K-) um mittels empirischer Quantilfunktion ein Intervall für den cut-off x_p (liefert die Spezifität p) zu ermitteln: $[X_p, X^p]$. Im zweiten Schritt werden unter Verwendung der Daten der Kranken (K+) die Sensitivitäten zu den cut-offs X_p und X^p ermittelt: $[Sens(X_p), Sens(X^p)]$. Damit wird dann das Konfidenzintervall berechnet: $Sens(X^p) - z * \sqrt{Sens(X^p) * (1 - Sens(X^p)) / n}$, mit n Kranken und entsprechendem Quantil z , liefert die untere Grenze.

Programmierung:

Die Umsetzung mit SAS verwendet nur SAS/BASE[®] und zwar im Kern ein Aufruf von proc sql. Im ersten Schritt wird ein Konfidenzintervall für die Spezifität ermittelt (`_plzphi`, `_puzphi`), für dessen Grenzen entsprechende cut-offs X_p und X^p zu finden sind: finde den cut-off, so dass dort die Spezifität minimal aber nicht kleiner ist als `_plzphi` (bzw. `_puzphi`):

```
proc sql;
  create table __ci4 as
  select b.*, a.&cutoff. as _xpl, a.&spec. as _tmpsl
  from __ci3 as b left join __spec as a on 1
  where (a.&spec. ge b._plzphi)
  group by b.&cutoff.
  having _tmpsl = min(_tmpsl);
  [...]
quit;
```

Der Datensatz a enthält die cut-offs, die unterschiedliche Spezifitäten liefern (Daten der Gesunden). Das gleiche wird für X^p (`_xpu`) gemacht und der zweite Schritt sieht fast genauso aus.

Anwendung:

Das Macro %roccih (confidence intervals for ROC using Hilgers method) ist einfach anwendbar und für die Berechnung für *ein* Testverfahren ausgelegt. Wenn *mehrere* Testverfahren vorliegen und es sollen *mehrere* Konfidenzbänder berechnet werden, muss das Macro mehrfach aufgerufen werden. Durch den Parameter prefix können die Ausgabevariablen mit einem prefix versehen werden. Weitere Übergabeparameter sind: data (Datensatz), alpha (Niveau), cutoff (Variable die den cut-off enthält), sens (Variable mit geschätzter Sensitivität), spec (Variable, mit geschätzter Spezifität), noe (Variable, mit der Anzahl der events=Kranken), n (Variable mit der Gesamtanzahl), out (Name des Ausgabe-Datensatzes), print (Schlüsselwort: yes/no, Ergebnis wird gedruckt), clean (Schlüsselwort: yes/no, temporäre Dateien werden gelöscht), help (Schlüsselwort: yes/no, Parameter mit Erklärungen werden ins log-Fenster gedruckt, ohne dass der Konfidenzbereich berechnet wird).

```
%macro roccih(data=_last_, alpha=0.05, cutoff=_cutoff1,
              sens=_sens, spec=_spec, noe=_events, n=_n,
              out=result, prefix=, print=no, clean=yes, help = no);
```

Um die Anwendung noch einfacher zu gestalten, wird das Macro intern in %rocn aufgerufen. %rocn (für mehrere Testverfahren) verwendet %rocl (für ein Testverfahren), %pv (Ermittlung von Vorhersagewerten, ggf. prävalenzabhängig), %cibinomexact (Berechnung von exakten Konfidenzintervallen), %roccih (Konfidenzband nach Hilgers).

```
%macro rocn(inset=_last_, tests=, event=, alpha=0.05, prev=,
            outset1=, outset2=result, se1sp0=yes, se0sp1=yes,
            delsens=no, delspec=no, lformat=best.,
            testpos=, clean=yes, help=no, options=nonotes);
```

Das Macro %rocn hat vor allem die Eingabeparameter inset (Name des Datensatzes), tests (Variablen zur Spezifikation der verschiedenen Testverfahren), event (Variable für die binäre Zielgröße), alpha (Niveau) und im wesentlichen den Ausgabeparameter outset2 (Name des Ausgabedatensatzes). Die anderen Parameter spielen eine untergeordnete Rolle. Eine Dokumentation befindet sich im Macro selbst. Geliefert werden viele Variablen. Beschrieben wurden bereits (Witte 2003):

```
_test, _fn, _cn, _cp, _fp, _n, _negative, _positive, _events, _nonevents,
_cutoff1, _cutoff2, _cutoff1_label, _cutoff2_label, _lmspec, _youden
```

Neu sind die Ausgabevariablen für Prävalenz $P(K+)$, prädiktive Werte $P(K+ | T+)$ und $P(K- | T-)$ und Trefferrate jeweils mit punktweisen Konfidenzintervallen:

```
_prev, _npv, _ppv, _hits, _acc,
_sens_exlcl _sens_exucl _spec_exlcl _spec_exucl
_acc_exlcl _acc_exucl _ppv_exlcl _ppv_exucl _npv_exlcl _npv_exucl
```

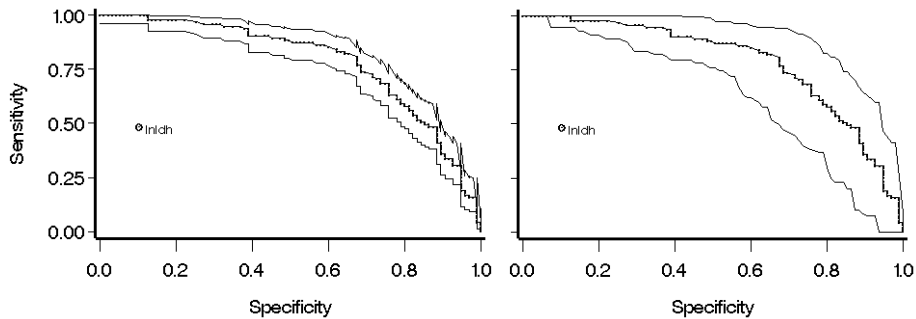
Werden externe Prävalenzen vorgegeben, wird für jede Prävalenz ein Satz von Variablen erzeugt, jeweils mit der Prävalenz im Variablennamen. Von wesentlichem Interesse für das Konfidenzband sind aber nur:

```
_sens, _spec, _lclh, _uclh
```

Die graphische Darstellung kann mit der Prozedur gplot umgesetzt werden, dann steht die ganze Funktionalität von SAS/GRAPH[®] zur Verfügung. Ein spezielles, für diese Anwendung geschriebenes, Macro %rocplot liefert die nötigen Graphiken. Die Parameter lcl und ucl können die Variablen für die untere bzw. obere Konfidenzintervallgrenze enthalten. So liefert das schon oben verwendete Beispiel von Schäfer mit

```
filename ausgabe 'ldh-line2.pdf';
%rocplot(gfile=ausgabe, gdev=pdf, inset=lnldh2, by=_test,
         lcl=_sens_exlcl, ucl=_sens_exucl);
```

die Abbildung 4. Punktweise Konfidenzintervalle sind im Allgemeinen nicht zu empfehlen, weil nur eine der Richtungen (hier Sensitivität) abhängig vom cutoff betrachtet wird.



Abbildungen 4 und 5: Punktweise Konfidenzintervalle für die Sensitivität bzw. punktweise Konfidenzbänder nach Hilgers.

Und die Abbildung 5 mit einem Konfidenzband wird erzeugt durch

```
filename ausgabe 'ldh-line3.pdf';
%rocplot(gfile=ausgabe, gdev=pdf, inset=ldh2, by=_test,
         lcl=_lclh, ucl=_uclh);
```

4 Beispiele

Um die graphischen Möglichkeiten darzustellen, wurde ein Datensatz (random) mit 400 Beobachtungen generiert, der drei verschiedene Testverfahren beinhaltet, die auf der Normalverteilung beruhen.

```
%rocn(inset=random, event=diag, tests=RAN1 RAN2 RAN3, outset2=random2);
```

Dieser Aufruf liefert die notwendigen Variablen, um die Konfidenzbänder darzustellen. Der Aufruf von %rocplot sieht immer gleich aus, es werden lediglich die Parameter cotype (Schlüsselwort: line oder area, Darstellungsform) und inset (Name des Datensatzes und Auswahl der Testverfahren) verändert, um die Abbildungen 6-11 zu erhalten.


```
%rocplot(gfile=ausgabe, gdev=pdf, by=_test, lcl=_lclh, ucl=_uclh,
citype=line, inset=random2(where=( _test in("RAN1", "RAN2", "RAN3"))));
```

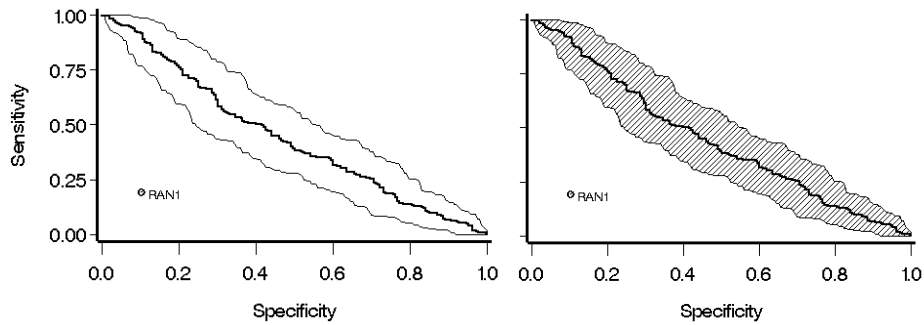


Abbildung 6 und 7: Ein Konfidenzbänder mit citype=line bzw. citype=area

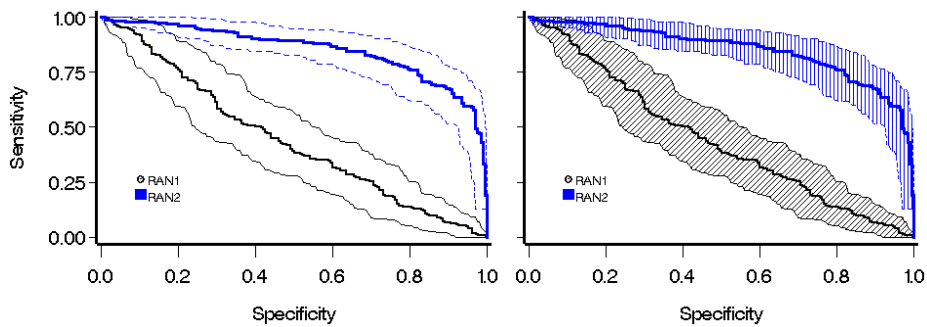


Abbildung 8 und 9: Zwei Konfidenzbänder mit citype=line bzw. citype=area

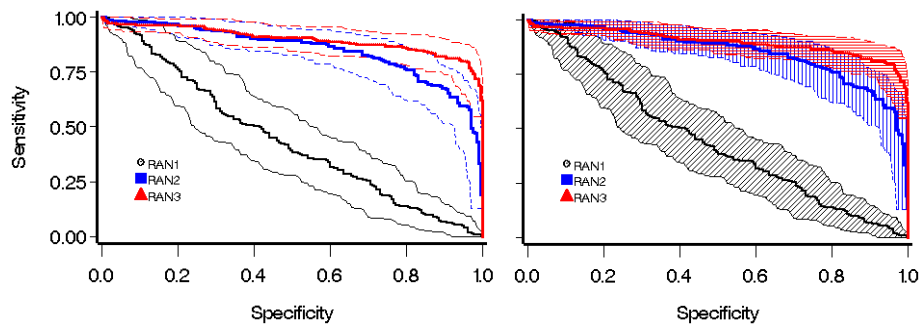


Abbildung 10 und 11: Drei Konfidenzbänder mit citype=line bzw. citype=area

Mit dem Macro %rocplot können bis zu 4 Testverfahren und deren Konfidenzbänder standardmäßig dargestellt werden. Bei mehr Verfahren müssen die entsprechenden symbol-statements selbst geschrieben werden. Ein Annotate-Datensatz kann mit %rocn einfach erstellt (outset2=out anno) und später in %rocplot verwendet werden (Witte 2003).

5 Diskussion

ROC-Kurve werden häufig dargestellt, so dass auch SAS-Anwendungen funktionsfähig sein sollten. Zusätzlich zu der ROC-Kurve ist ein Konfidenzband ein wichtiger Schritt zur Interpretation der Kurve, denn für eine vorgegebene Spezifität kann das Konfidenzintervall für die Sensitivität abgelesen werden. Die Methode nach Hilgers ist im Macro %roccih umgesetzt und in %rocn integriert, so dass die Anwendung sehr einfach ist. %rocplot liefert dann Standardgraphiken, die direkt für Publikationen verwendet werden können. Es sei darauf hingewiesen, dass die Konfidenzbänder keinen statistischen Test zum Vergleich von ROC-Kurven liefern. Dabei üblich ist der Vergleich der Flächen unter den ROC-Kurven (z.B. %auroc von Jochem König). Die Umsetzung von weiteren (effizienteren und auch simultanen) Konfidenzbändern wäre sinnvoll.

Die Programme (incl. Beispielprogramm) sind als Freeware erhältlich, bitte schreiben Sie bei Interesse eine e-mail an sas@imbi.uni-heidelberg.de.

Danksagung:

Für die Mithilfe bei der Umsetzung der Darstellung von Konfidenzbändern mit schraffierten Flächen möchten wir Christina Klose recht herzlich danken.

Literatur

1. Campbell, G. (1994) Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in Medicine*, **13**, 499-508
2. Daly, L. (1992) Simple SAS macro for the calculation of exact binomial and Poisson confidence limits. *Comput. Biol. Med.*, **22**, 351-361
3. Hilgers, R.A. (1991) Distribution-free confidence bounds for ROC curves. *Meth. of Inform. in Med.*, **30**, 96-101
4. Jensen, K., Müller, H.H. und Schäfer H. (2000). Regional confidence bands for ROC curves. *Statistics in Medicine*, **19**, 493:509
5. Schäfer, H. (1994). Efficient confidence bounds for ROC curves. *Statistics in Medicine*, **13**, 1551-1561
6. Witte, S. (2003) Macro zur Erstellung von ROC-Kurven. Proceedingsbeitrag zur 6. KSFE, Universität Dortmund, 375-386