臨床検査診断能のロバストな比較：ROC の AUC を中心に
Evaluation of The Performances of Diagnostic Tests Estimation of Areas Under ROC curves by Maximum Likelihood and Bootstrap Method, etc

古川敏仁
株式会社バイオスタティスティカル　リサーチ

1. 概要

本報告書は 1997 年 10 月 1 日　東京大学医学部　疫学・生物学教室で発表した原稿の修正版である。臨床検査評価のための中級・上級者を対象としたテキストである。

日常臨床や臨床試験などでは臨床検査は重要な位置を占めているが、そこで使われている検査は、必ずしもその有用性がきちんと評価されているわけではない。そこで、今回は最も簡単なシチュエーション：

患者が２つの状態、Ａ：対象疾患有りと、Ｂ：対象疾患無し

それを検査値Ｘ1、Ｘ2・・・・などで疾患を単独で診断する場合、

どの検査が優れているかを、ROC 曲線などのような検査診断能のロバストな比較方法について考えてみる。

2. 従来の Index による評価法

正常値Ｔを用いて、検査ｘの値に対し

ｘ＞Ｔ：患者の検査値がＴを越えると検査結果はＡ：陽性（有病と推定）

ｘ≦Ｔ：　〃　がＴ以下ならば　〃　Ｂ：陰性（無病と推定）

と検査結果を利用する場合が多い。

今、下記のように用語を定義しよう。

True Positive：有病者が検査陽性である場合。

True Negative：無病者が検査陰性である場合。

False Positive：無病者が検査陽性である場合。

False Negative：有病者が検査陰性である場合。

この場合、従来は、患者の真の状態（より正確度の高い情報をもとにした判定）と検査結果を比較して、以下のような指標で検査の能力を比較している。

(1)正確度（Accuracy) :

$$Accuracy = \frac{\text{True Positive} + \text{True Negative}}{\text{全例数}}$$

(2)感度（Sensitivity）　　ＦＮＦ：False-Negative Fraction

１－感度

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{有病例数}}$$

(3)特異度（Specificity）　　ＦＰＦ：False-Positive Fraction

１－特異度

$$\text{Specificity} = \frac{\text{True Negative}}{\text{無病例数}}$$

(4)陽性的中度（ＰＰ）：Positive Predictability

$$\text{Positive Predictability} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

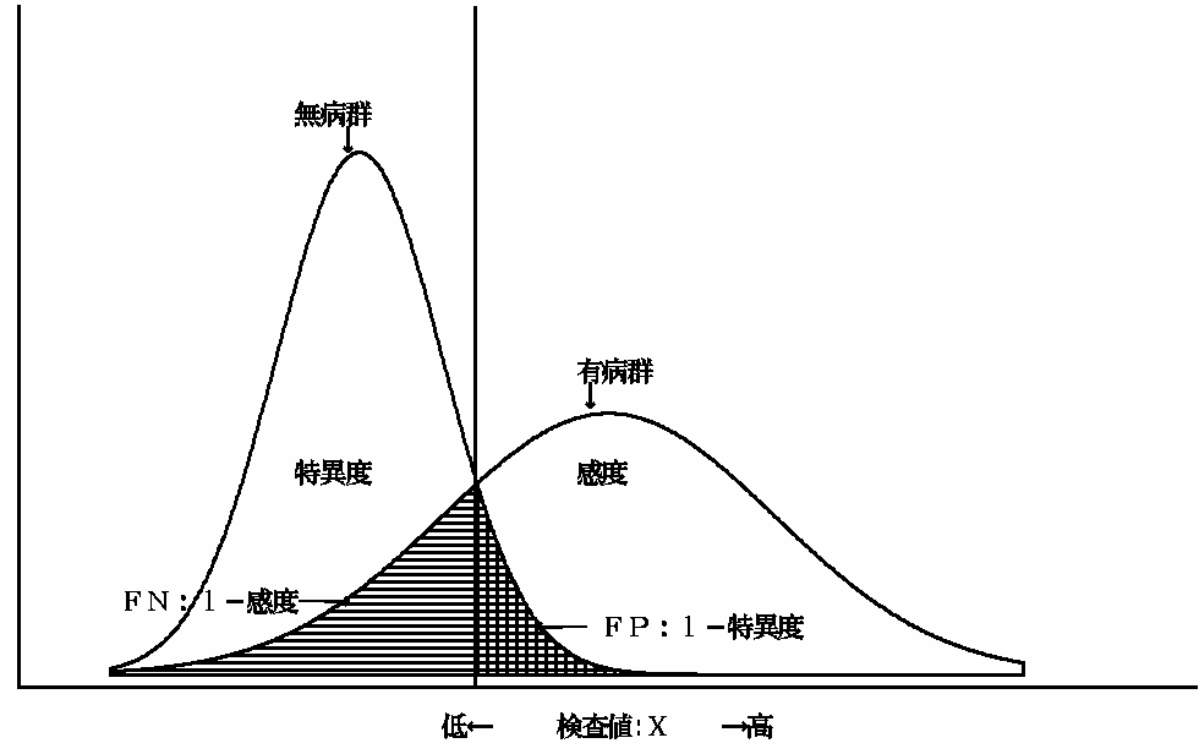
(5)陰性的中度（ＮＰ）：Negative Predictability

$$\text{Negative Predictability} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Negative}}$$

Fig.1

閾値Ｔ

有病群と無病群それぞれの検査Ｘ値上での分布を、ある閾値Ｔで区分すると、
感度、特異度、FPF、FNF の４つの領域に区分できる。



しかしながら、これらの指標による評価では、以下の２つの問題がある。
a)有病率（prevalence)の影響
b)閾値の設定について

a)有病率（prevalence)の影響
検査を評価すための一見妥当な指標に正確度がある。これは、検査的中例を全症例で除したものであるが、例えば疾患の有病率が 1%である場合、閾値をその検査の最大値を越えるところに設定する（検査結果を全て陰性と判定する）と、全く無意味な検査でも正確度は 99%である（(1)参照)。陽性的中率，陰性的中率も同様に有病率の影響を受け、陽性的中率は、同じ感度、特異度の検査であっても有病率が低いほど低くなる（(2)参照)。

(1)　感度 0%、特異度 100%　閾値∞
　無意味な検査の場合のパラドックス

	有病率		
	50%	10%	1%
正確度	50%	90%	99%
陽性的中度	0%	0%	0%

(2)　感度 90%、特異度 90%
非常によい検査の場合のパラドックス

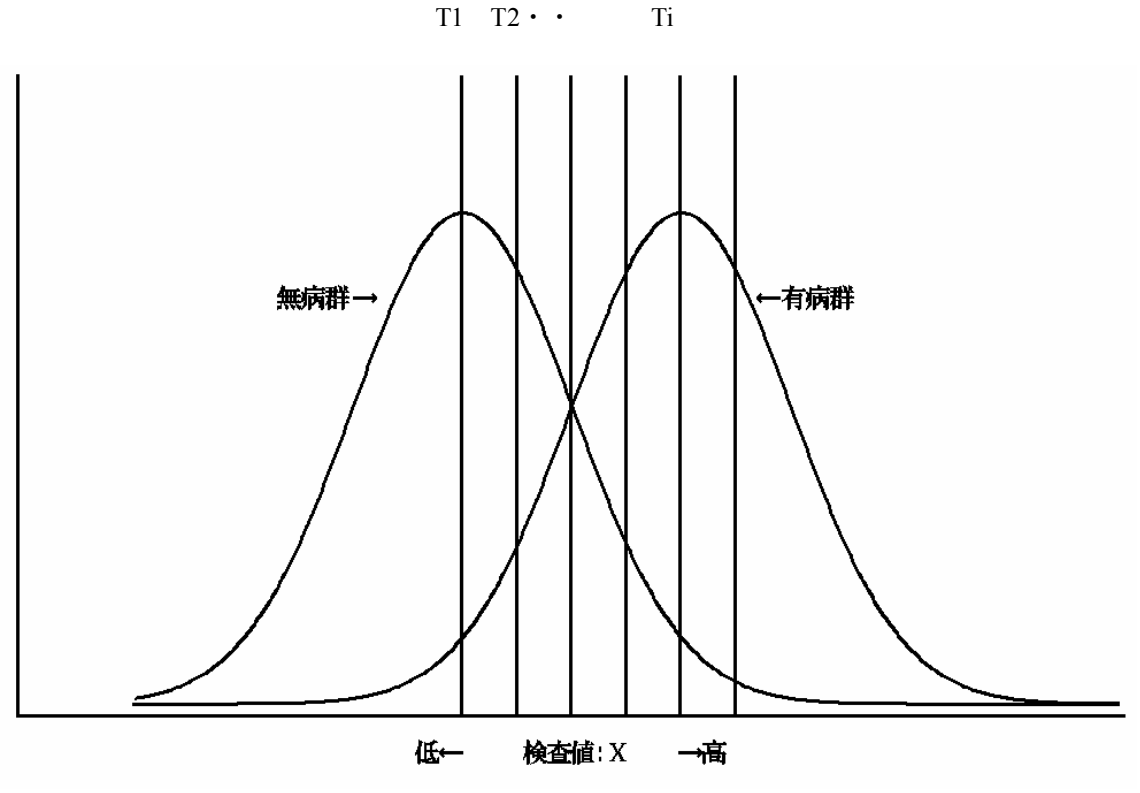
	有病率		
	50%	10%	1%
正確度	90%	90%	90 %
陽性的中度	90%	50%	8.3%

ｂ)閾値の設定について　感度、特異度のような指標は有病率の影響を受けないが、FIG.2 のように閾値を変化させれば、(1)～(5)全ての指標は閾値に応じて違った値となる。
　また、一般的に閾値は、検査X値上の無病群、有病群の分布の確率密度関数 f(x),g(x)と、有病群検出の重みWから、(6)式で示される L(x)で求めることができる。

$$L(x) = \frac{g(x) \bullet w}{f(x)} \qquad (6)$$

しかし、標準的な有病率や、重みWを決定するための疾患の重要性を決めることは難しく、従って検査を評価する場合、一つの閾値での判定結果には問題が多すぎる。

FIG.2　閾値の変化と区分領域の変化



3. 検査診断能の定義とその評価方法

実際の臨床への検査診断能の評価は、有病率、感度と特異度どちらを重用視するかなどの情報を考慮した上で決定すべきものであるが、検査診断能の比較評価、すなわち、有病群と無病群を判別するという検査の基本的特性の比較に関して言えば、有病群と無病群の例数を 1：1、感度と特異度の重要性の重みを等しい、すなわち(6)式の w=1 という条件で評価するのが、スクリーニング的な検査性能の比較には適している。

これら検査診断能を比較する方法に下記の 3 手段が考えられる。

1. 有病群と無病群の検査値順位のずれを、検査値間で比較する方法
2. ROC 分析（Relative もしくは Receiver Operating Characteristic curve）
3. ロジスティック回帰分析、判別分析のようなモデル手法

3.1. 有病群と無病群の検査値順位のずれを、検査値間で比較する方法（順位平均比較法）

一般的に性能の良い検査ほど検査値上の分布で有病群と無病群は分かれていくはずである（鑑別されていく）。この、鑑別能は、有病群と無病群の順位平均の差となって観測される。そこで、順位平均の差を、順位を応答変数、検査項目、疾病有無、検査項目と疾病有無の交互作用を応答変数とした線型モデルで、検査項目と疾病有無の交互作用の統計学的有意性を検討すれば、検査鑑別能の直接的な比較ができる（付録 1. SAS プログラム例参照）。

この手法のメリットは下記である。

1) 検査鑑別能の比較が検定ベースで可能となる。
2) ノンパラメトリックな手法であるため、検査値の分布に依存しない。

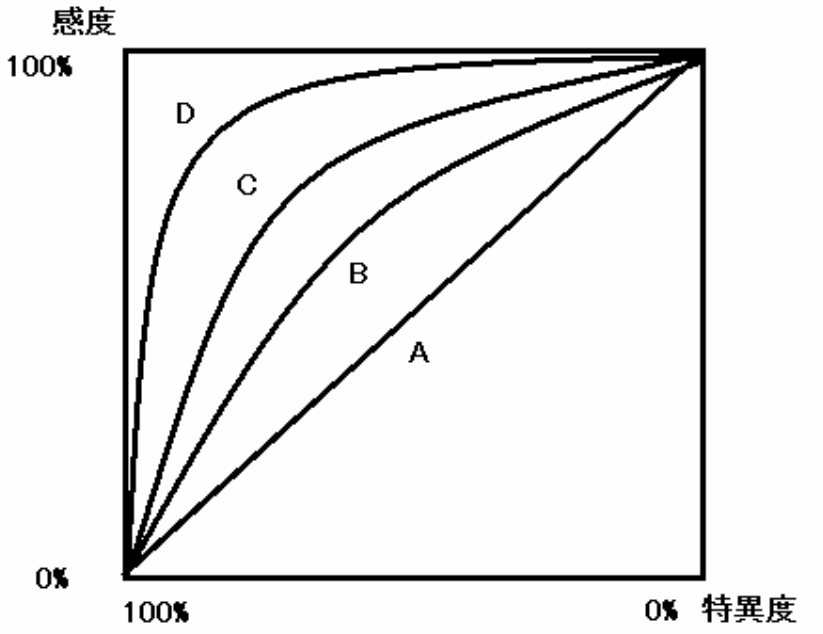
問題点は、検定以外の利用できる情報がないため、評価例数等の条件に注意しないと、同時に測定された検査以外は比較できないことである。

3.2.　ROC 分析（Relative もしくは Receiver Operating Characteristic curve）

A) 曲線からの検査性能の判断

ROC 曲線は、閾値を変化させ、それぞれの閾値での感度を縦軸に、ＦＰＦ（偽陽性：１－特異度）を横軸にプロットしたものである（図１）。　ROC 曲線では、全く鑑別のない検査は、A のような対角線上に曲線を描き、鑑別能が向上するほど、B,C,D のように、対角線から左上に弧を引く曲線となり、鑑別能 100%の検査は、左辺-上辺上の曲線となる。

図1．　ROC曲線



B) 曲線の接線の傾き－閾値の適切な設定

ROC 曲線の接線の傾きは、その地点での検査値 x での有病、無病群の分布の尤度比 L(x)　{(6)式における W=1 の場合}を示している（図 2)。

$$x\text{検査値での尤度比}\quad L(x) = \frac{g(x)}{f(x)}$$

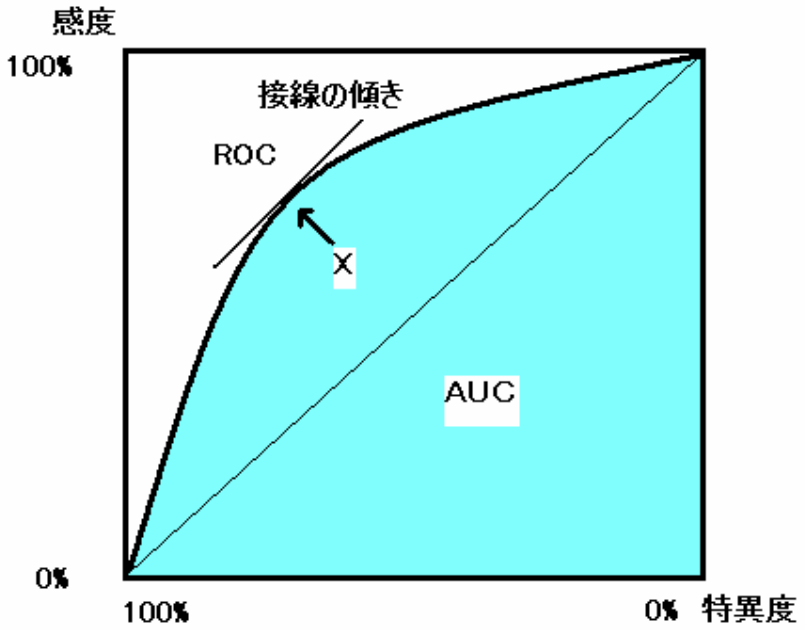
ここから、逆に。各疾患ごとに閾値として臨床的に妥当な L(閾値)を設定して、ROC 曲線からそれに該当する接線の傾きを持つ検査値を推定すれば、それが適切な閾値となる。

例えば、L(閾値)＝1.0　有病率が１の場合、両群を同じ重みで区分したい。

0.5　　〃　　有病群を無病群の２倍の重みで検出したい。

ROC 曲線の接線の傾きは、例えば横軸の区間幅を約 5%に設定した移動平均的な手法により簡単に計算できる。また、信頼区間は Bootstrap 法などから求めることができる（付録 2、付録 4)。

図2　ROCのAUCと接線



C)　ＡＵＣ－平均的な正確度の指標

AUC(Area Under the Curve)[2]はROC曲線の曲線下面積のことであり、直感的にはROC曲線の左上角への近づき程度をしめす指標である。また、数学的には平均的な検査の正確度を示している。

仮に台形公式でＡＵＣを計算する事を考えてみよう。

閾値がＴ1，T2・・Ｔi・・・Ｔmとし
　　Ｔi～Ｔｉ－１に入る無病群の例数をＮｉ　無病群の例数をｎｎ
　　ｘ＞Ｔｉの領域の有病群の例数をＤｉ　有病群の例数をｎｄ

とすれば　Ｔｉ－１～Ｔｉ区間に対応する横軸の幅は　Ｎｉ／ｎｎ
　　Ｔｉ－１～Ｔｉに対応する縦軸値は Ｄｉ－１/ｎｄ，Ｄｉ/ｎｄ

(7)　$$AUC=\sum_{i=1}^{m}\left\{\left(\frac{N_i}{n_N}\right)\left(\frac{D_{i-1}+D_i}{2n_D}\right)\right\}$$

(8)　$$AUC=\sum_{i=1}^{m}\left\{f(x_i)\left(\frac{D_{i-1}+D_i}{2n_D}\right)\right\}$$　　ただし、$f(x_i)=\frac{N_i}{n_N}$

すると AUC は(7)、(8)式のように展開でき、AUC は f(x)の密度関数で分布する無病群の感度の期待値、もしくは g(x)で分布する有病群に対する特異度の期待値になっていることが分かる。

台形法による AUC の期待値、分散値は、MULTTEST Procedure の Bootstrap 法を利用して簡単に推定することができる（付録 2、付録 4)。

また、無病群、有病群の分布を仮定したパラメトリックな方法が Metz [1), 2), 3)] らによって、提案されている。

いま、有病群、無病群がそれぞれ、N（μN,σN２）、N（μD,σD２)に従うと仮定すると、a、b２つのパラメーターから、

（９）　$a = \frac{\lambda_D - \lambda_N}{\sigma_D}$　　$b = \frac{\sigma_N}{\sigma_D}$

標準正規分布の分布関数Fより、(10),(11)のように求められる。

（１０）　$AUC = F\left(\frac{a}{\sqrt{1+b^2}}\right)$

（１１）　$$Var(AUC) = \left[\frac{1}{\sqrt{2\pi(1+b^2)}}\exp\left\{-\frac{1}{2}\left(\frac{a}{\sqrt{1+b^2}}\right)^2\right\}\right]Var(a)$$
$$+\left[-\frac{ab}{\sqrt{2\pi}}\frac{1}{(1+b^2)\sqrt{1+b^2}}\exp\left\{-\frac{1}{2}\left(\frac{a}{\sqrt{1+b^2}}\right)^2\right\}\right]Var(b)$$
$$+2\left[\frac{1}{\sqrt{2\pi(1+b^2)}}\exp\left\{-\frac{1}{2}\left(\frac{a}{\sqrt{1+b^2}}\right)^2\right\}\right]\times$$
$$\left[-\frac{ab}{\sqrt{2\pi}}\frac{1}{(1+b^2)\sqrt{1+b^2}}\exp\left\{-\frac{1}{2}\left(\frac{a}{\sqrt{1+b^2}}\right)^2\right\}\right]Cov(a,b)$$

Metz らは閾値をおおよそ 10 回変化させ、約 10 カテゴリー区分に属する無病群と有病群の数を基にした最尤法による a,b の推定方法を提案している。
また、かれらはそのカテゴリー区分方式のよりロバストな台形法による AUC の推定法も提案している。

ＸＤｊ　ｊ番目のカテゴリーに属する有病群の例数
ＸＮｊ　ｊ番目のカテゴリーに属する有病群の例数

（１２）　$$\hat{AUC} = \{\sum_{j=1}^{n}(x_{Nj}(\sum_{i>j}x_{Di}))+\frac{1}{2}\sum_{j=1}^{n}(x_{Nj}x_{Dj})\}/(N_N\cdot N_D)$$

$$Q_1=\sum_{j=1}^{n}\left[x_{Nj}\left\{(\sum_{i>j}x_{Di})^2+(\sum_{i>j}x_{Di})x_{Dj}+\frac{1}{3}x_{Dj}^2\right\}\right]/(N_N\cdot N_D^2)$$

$$Q_2=\sum_{j=1}^{n}\left[x_{Dj}\left\{\left(\sum_{i<j}x_{Ni}\right)^2+\left(\sum_{i<j}x_{Ni}\right)x_{Nj}+\frac{1}{3}x_{Nj}^2\right\}\right]/(N_D\cdot N_N^2)$$

（１３）　$$Var(\hat{AUC})=\frac{\hat{AUC}(1-\hat{AUC})+(N_D-1)(Q_1-\hat{AUC}^2)+(N_N-1)(Q_2-\hat{AUC}^2)}{N_D\cdot N_N}$$

4. 結果・結論

4.1.　ＡＵＣ計算結果の比較

台形法、Bootstrap 法、Metz の２法によるＡＵＣ、標準偏差の比較を行った（表１）。

例：卵巣癌 70 例、良性卵巣腫瘍 251 例による比較

	ＡＵＣ	標準偏差
台形法単純計算	0.8478	－－－
Bootstrap 台形法 N=500	0.8479	0.02778
最尤法（パラメトリック） 10 カテゴリー	0.8451	0.02913
最尤法（台形法　　）	0.8431	0.02905

いずれの、方法も例数が多いときは大差がないが、

- Metz のパラメトリック法は分布が歪んでいるときには計算不能となる。
- Metz のロバスト法は例数が少ないときは計算不能となる。

ただし、この標準偏差は、AUC を確率値としたときの２項分布からの推定値とほぼ同じであり、理論的にも整合性が取れている。

4.2.　２検査項目間の比較

（順位平均比較法）と、上記で求めた AUC 標準偏差から、下記の式で検定した結果を比較したところ、両者の検定結果はほとんど同じであった。

（１４）　$$Z_{AUC}=\frac{(AUC_1-AUC_2)}{\sqrt{(S_1^2/N_1+S_2^2/N_2)}}$$

ROC 曲線の AUC の比較は、基本的には判別分析（SAS：DISCRIM プロシジャ）やロジスティック回帰分析（LOGISTIC　プロシジャ）による検査変数の有意性の検討と基本的には同じである[(4)]。以下の特性を考慮し、ROC 解析とモデル解析、あるいは順位平均のノンパラメトリック検定を組合せ、検査特性の全体像を検討することが重要である。

検査診断能の評価におけるモデル解析の利点

1)　過去の検査データから、疾患の有無の確率を推定する統計モデルの作成が可能。

2) 他の変数の影響を除外（調整）したり、逆に、他の変数との組合せ効果などのように、検査項目の組合せ効果も検討できる。

検査診断能の評価におけるモデル解析時の制約

1) データの分布がモデルのロバスト性に影響する。
例：線形判別分析　正規性、等分散性などの検討が必要
2) 小数データからのモデルは、そのデータに依存する。すなわち、モデルの普遍性を別のデータなどで確認する必要がある。

5. まとめ

1. ROC 解析は検査の区分能を解析する手法であって、臨床的有用性を解析する手法ではない。
2. ROC 解析は直感的に分かりやすく、数学的にも意義が深い。
正確度の期待値が AUC となる。
接線の傾きは、有病群、無病群の確率密度関数の尤度比
3. ROC 曲線の AUC 推定に関して、３つの方法の SAS モジュールを作成した。
4. AUC の分散推定には MULTTEST Procedure の Bootstrap 法が利用できる。
5. AUC の推定値とその分散による、検査の有効性の検定は平均順位の検定（Wilcoxon 検定）と良く一致している。
6. ROC 解析と Logistic 解析のような手法を同時に用いることは有用である。特に、モデルベースドな手法により、多変量の情報の有効性を吟味し、ROC で、実際的な意味を確認することは重要である。

参考文献

1) Metz CE: Statistical analysis of ROC data in evaluating diagnostic performance. In:Multiple regression analysis:Applications in the health science.(Herbert D.,Myer R.,eds.). American Institute of physics, New York, 52-56, (1986)
2) Hanley JA and McNeil BJ: The Meaning and Use of the Area under a Receiver Operating Characteristic(ROC)Curve. Radiology, Vol.143 , 29-36(1982)
3) Dorfman DD, Alf E: Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data. Journal of Mathematical Psychology ,6,487-496(1969)
4) Grey DR, Morgan BJT: Some aspects of ROC curve-fitting: Normal and logistic models. Journal of Mathematical Psychology, 9,128-139(1972)

付録 1.　有病群と無病群の検査値順位のずれを、検査値間で比較する SAS プログラム例

```
/* 対応する検査項目　X1、X2 の、疾患「あり」、「なし」（変数　Diag）を含めた順位を作成 */
PROC RANK DATA=I OUT=Y1 ;
        VAR X1 ;
        RANKS RK;
   RUN;
   PROC RANK DATA= OUT=Y2 ;
        VAR X2 ;
        RANKS RK;
   RUN;
   DATA Y1;
      SET Y1;
         TEST='X1';
   RUN;

   DATA Y2;
      SET Y2;
         TEST='X2';
   RUN;
   DATA ANL;
    SET Y1 Y2;
   RUN;

   /* 検査　TEST と Diag の交互作用を検定　　*/
   PROC GLM;
    CLASS L TEST;
       MODEL RK=DIAG TEST DIAG*TEST;
   RUN;
```

付録 2. FREQ プロシジャを利用した ROC 曲線作図

```
/* GRP =1   疾患あり：有病群
   GRP=0   疾患なし：無病群
   KENSA：検査値
*/
PROC SORT DATA=TEST;     BY GRP;     RUN;

PROC FREQ DATA=TEST;
   BY GRP;
   TABLES KENSA/ OUT=OUT NOPRINT;
 RUN;

PROC SORT DATA=OUT;   BY KENSA;      RUN;

DATA out;
   RETAIN SEN PID 100 SP NIH KN 0;
   SET out;
        IF GRP = 1 THEN SEN=SEN-PERCENT;
        IF GRP = 0 THEN   SP=SP+PERCENT;
      _SENSIT_=SEN;     NIH=SP;
      _1MSPEC_=100-NIH;
      RINJI=PID+_1MSPEC_;
   KEEP KENSA _SENSIT_   RINJI _1MSPEC_;
 RUN;

PROC SORT DATA =OUT;     BY KENSA RINJI;      RUN;

DATA OUT;              /*  閾値ごとの整理   */
   RETAIN AA 0;
   SET OUT;
       IF KENSA=AA THEN DELETE ;
      AA=KENSA;
   KEEP KENSA _SENSIT_ _1MSPEC_;
   RUN;

DATA OUTF;             /*   100，100％点の付加   */
   INPUT PID _1MSPEC_;
      CARDS;
         100 100   ;
   RUN;

DATA OUT;
   SET OUTF OUT;
 RUN;

PROC GPLOT data=ROC2;      /*   作図       */
   SYMBOL1 I=JOIN W=2 V=NONE C=black L=1;
      PLOT _SENSIT_*_1MSPEC_          / FRAME NOLEGEND
           VAXIS=AXIS1 HM=0 VM=0          HAXIS=AXIS2;
 RUN;
```

付録３　ロジスティックプロシジャを利用した ROC 曲線作図

```
/*   GRP：グループ変数   0:無病   1:有病
      KENNSA：ＲＯＣ分析対象項目の変数   */

PROC LOGISTIC DATA=TEST DESCENDING NOPRINT;
    MODEL CC1=KENSA / OUTROC=ROC1;
 RUN;
DATA ROC2;
    SET ROC1;
  _SENSIT_=_SENSIT_*100;        _1MSPEC_=_1MSPEC_*100;
 RUN;

PROC GPLOT data=ROC2;       /*   作図       */
  SYMBOL1 I=JOIN W=2 V=NONE C=black L=1;
     PLOT _SENSIT_*_1MSPEC_            / FRAME NOLEGEND
          VAXIS=AXIS1 HM=0 VM=0          HAXIS=AXIS2;
 RUN;
```

付録 4 MULTTEST を利用したＡＵＣの分散の推定

```
/* CLASS=1 500 回の復元抽出　*/
PROC SORT DATA=TEST;
  BY GRP;     RUN;
PROC MULTTEST NSAMPLE=500 DATA=TEST OUTSAMP=OUT SEED=12345 NOCENTER
    NOPRINT BOOTSTRAP;
     BY GRP;       TEST MEAN(KENSA);       CLASS CLASS;
  RUN;
%MACRO BUNKATU;
    %DO I=1 %TO 500;
        DATA OUT&I;       SET OUT;
            IF _SAMPLE_=&I;
     %END;
%MEND BUNKATU;
%BUNKATU
%MACRO AUC;
      %DO I=1 %TO 500;
          PROC FREQ DATA=OUT&I;
              TABLE KENSA / OUT=A NOPRINT;
              BY GRP;
          PROC SORT DATA=A;              BY KENSA;
           RUN;
  **データセット A に対する　PRG2　ルーチンの実施（省略）**

           DATA A2;　/*　A2 データの AUC の計算　*/
              RETAIN A B 100;
              SET A;
                 S=(A+_SENSIT_)/2*(B-_1MSPEC_)/10000;
                 A=_SENSIT_;          B=_1MSPEC_;
           PROC MEANS DATA=A2 SUM;
              VAR S;
              OUTPUT OUT=AUC SUM=AUC;
           DATA AUCALL;
              SET AUC AUCALL;         RUN;
     %END;
%MEND AUC;
%AUC
DATA AUCALL;
        SET AUCALL(FIRSTOBS=2);
PROC MEANS DATA=AUCALL MEAN STD;
        VAR AUC;
RUN;
```