相関構造をもつROC曲線に対する検出力の算出

Power analysis in correlated receiver operating characteristic (ROC) curves : a simulation study

西本　尚樹*　伊藤　陽一**　菊池　明泰*　佐々木　恒平*
奥山　豪*　織田　圭一*　真田　哲也*
板東　道夫*　早川　修*　北間　正崇*

Naoki Nishimoto, Yoichi M. Ito, Akihiro Kikuchi
Kohei Sasaki, Go Okuyama, Keiichi Oda,
Tetsuya Sanada, Michio Bandoh,
Osamu Hayakawa and Masataka Kitama

Abstract

The purpose of this study is to evaluate power when comparing ROC curves with multi-reader and multi-modality case (i.e. correlated data). NA Obuchowski (1995) proposed a methodology for sample size calculation in above situation. The intuitive understanding for the power is required for developing the clinical study design. We used Obuchowski's algorithm for sample size calculation. Simulation data were areas under the curve (AUC) of 0.848±0.059 for control and 0.883±0.050 for test, respectively. We computed statistical powers for the articles with a range of the AUC difference using SAS 9.3 (SAS Institute Inc.). We obtained the power for the range of 0.05 to 0.927. At the point of the difference of 0.035 in the article, power was 0.865. We could see that the power increased from 0.384 to 0.865 at the point of the article when the number of readers was increased. For the number of readers was 11, 12, the power was over 0.8 at the point of 0.035. The number of readers less than 10 produced the power less than 0.8 even if the difference of AUC would be greater than that in the article. We found that the statistical power has been changed when the number of readers was increased. We fixed the correlation between readers and modalities, thus, further investigation is required for the correlation parameter increase. (221 words)

1. はじめに

医療における検査の評価には、診断能の評価とともにROC(receiver operating characteristic)解析が用いられている。感度と1-特異度をプロットし、ROC 曲線を求めることで、全範囲の閾値に対する検査が陽性と判定される確からしさを描画するものである。放射線技術学領域では、(1)検査としての性能、(2)診断能としての二つの側面からROC 解析が使用されてきた[1-4]。特に、CAD(computer aided detection)の評価には、ROC曲線がよく用いられており、統計解析手法の選択とともに先行研究がなされている[5-8]。

しかしながら、実際に症例画像(case)とコントロール画像(control)を集める実験計画を立案するときに、何例の症例数を集積し、読影者(reader)に協力してもらうのがよいのか、一意に決定することは難しい。実験の計画において照度や読影の順序、ランダム化など、種々の条件を考慮せねばならず、multi-reader, multi-caseになる放射線技術学領域のROC実験は容易ではない。通常の症例数の設計とは異なり、ROC 実験の場合には読影者内の相関と同一症例内の相関を考慮しなければならない。解析手法にJackknife法を用いる場合も同様で、症例数が少な過ぎたり読影者が少な過ぎたりすれば、検出力(power, $1-\beta$)不足になり、新たに開発した手法と従来法の間に真の差があったとしても、差が検出できない可能性が生じる。また、あらかじめ計画された実験で、検出力不足であれば、被験者にとっては実験に協力しても統計的に意味のある結果が得られず、倫理的な問題も生じる。先行研究の症例数を見て、多ければ多いほど良いという論理のもと100例や200例で行っている研究も見られるが、症例数や検出力を計算していなければ、どの程度の確からしさで真の効果を推定できるのかが分からない。

しかしながら、multi-reader, multi-caseのROC

*北海道科学大学保健医療学部診療放射線学科
**北海道大学大学院医学研究科医学統計学分野

実験における症例数設計は複雑であり、読影者内の相関と同一症例内の相関を考慮するには、それぞれの効果をモデルに組み込んだ二元配置分散分析を用いた手法による症例数設計が必要である。初学者にとって、二元配置分散分析を用いた症例数の設計は、どの程度の効果があれば症例数がどういう傾向を持って増加するのかを直感的につかみにくい。従って、症例数の設計式から、グラフを描画することによって可視化を試みるが、条件が複数あるため、二次元的に考える必要がある。

2.目的

本研究の目的は、同一読影者、同一症例内に発生する相関を考慮する必要のある ROC 解析を実施する際における読影者の人数と検出力の関係を明らかにすることとした。

3.方法

3.1 ROC 曲線下面積のモデル

パラメトリックモデルを使って ROC 曲線下面積(AUC, Area under the curve)の検定を行う場合を想定し、検出力曲線を描画して、症例数の増加を視覚的・定量的に求めた。

i 番目の放射線診断機器(モダリティ)を用いて、j 番目の読影者が画像を k 回目に読影するときの AUC を θ_{ijk} として、以下のように記述する。

$$\theta_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

ここで、μ は AUC の全平均を表し、α_i は i=1,…,I のうち i 番目のモダリティの固定効果とする。β_j は j=1,…,J のうち j 番目の読影者の変量効果とし、平均 0、分散 σ_b^2 の正規分布に従うものとする。$(\alpha\beta)_{ij}$ は、平均 0、分散 σ_{ab}^2 に従う読影者とモダリティの交互作用項とする。これはすなわち、特定の読影者がある得意なモダリティによって読影能力や疾患の検出能力が上がることを仮定している。ただし、特定の読影者の能力が平均的に高いことには着目しないため、読影者の効果は変量効果とした。

このモデル化は、変量効果と固定効果が混合されているため、一般線形混合効果モデルとして広く知られている。分散分析・重回帰分析は一般的に、固定効果のみで用いられることが多いが、分散分析・重回帰分析のモデルを拡張したものが一般線形混合効果モデルに当たる。昨今では、薬学や医学などで、経時データの解析に用いられている[9]。

3.2 検出力の算出式

症例数を決定する上で、一般的には有意水準 α と検出力 $1-\beta$ を決定して求める事が多い。また、実際の ROC 解析を用いた研究に必要なものは症例数であるが、放射線技術学領域における研究には、収集可能な症例数の限界があることが多い。そこで、実施可能数として症例数を 30 例に固定し、得られる検出力をグラフにて描画した。検出力には以下の算出式を用いた。

まず、対立仮説の分布の非心パラメータを導く。対立仮説の分布は、帰無仮説に比べて幾分偏りが生じており、その偏り具合は非心パラメータを用いて表される。

$$\lambda = \frac{\frac{k}{2}(AUC_{control} - AUC_{test} - \delta)^2}{\sigma_{ab}^2 + \sigma_w^2 + \sigma_c^2\{(1-r_1) + (k-1)(r_2 - r_3)\}}$$

ここで、$AUC_{control}$ と AUC_{test} は、それぞれ対立仮説のコントロールの AUC とテスト(新規モダリティ)の AUC を表す。δ は 0 に設定する。その他の記号については、以下とする。

σ_w^2：読影者内の分散

σ_c^2：被験者間の分散

r_1：同一読影者内で、異なる放射線診断機器の相関

r_2：異なる読影者間で、同一の放射線診断機器における相関

r_3：異なる読影者間で、異なる放射線診断機器の相関

F 分布の逆関数 $F^{-1}(p)$ を用いて、有意水準 5%の F 値 f*を算出した。上記で求めた非心パラメータ λ を用いて、検出力 Power を求めた。

$$\text{Power} = 1 - F(f^*, df_{num}, df_{denom}, \lambda)$$

ここで、F(x)は F 分布の累積分布関数を表す。df_{num} と df_{denom} は、分子と分母の自由度を表す。

3.3 検出力計算の条件

コントロールとテストの 2 群を比較する研究を仮定し、被験者数を 30 例に固定した。Oda らは先行研究で、小肺結節の検出を行うために肋骨の抑制処理の有無について、12 名の読影者で評価を行った。コントロールの AUC は 0.848±0.059、肋骨の抑制処理を行った結果では 0.883±0.050 となり、診断の向上を報告した[10]。Oda らの研究を参考に、本研究ではベースラインの AUC を 0.847 と

し、0.001 ずつ新規の放射線診断機器群の AUC を増加させた。また、読影者の人数も 5 人から 12 人へ 1 人ずつ増加させて検出力の変化を描画した。同様に先行研究を参考にして、$\sigma_{ab}^2 = 0.000187$, $\sigma_w^2 = 0.0001, \sigma_c^2 = 0.0004$, r_1～r_3のそれぞれの相関を 0.1 に設定した。検出力の計算には SAS9.3(SAS Institute Inc.)を用いた。

4.結果

テストの AUC を X 軸に取り、検出力を Y 軸にプロットした図を示した。グラフからコントロールとテストの AUC の差が小さいときに、その差が大きくなるに従って、検出力は単調増加関数として増加する関係性が明らかになった。また、読影者を増やすことで、検出力を直線的に増加させることが可能である。しかしながら、読影者を 10 名以上に増やしても、検出力が 0.8 を超えたところから大きく増加はせず、検出力はプラトーに達した。

コントロールの AUC を 0.848、テストの AUC を 0.883（差 0.035）で固定したところ、評価者を 5 人から 12 人に増やした場合に、検出力は 0.384 から 0.865 まで上昇した。読影者が 10 人以下では AUC の差が小さい場合に、十分な検出力を確保できないが、読影者を増やすことで、検出力を高めることができる。

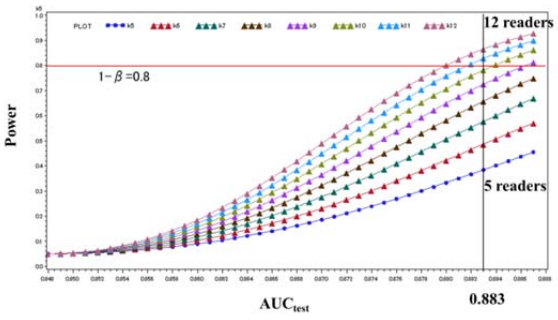

図 1 検出力と新規の放射線診断機器の AUC の関係。コントロールとの AUC の差が開くと、指数関数的に検出力が上昇する。

5.考察

検出力は、真に差があるものを検定で正しく差があると結論できる確率を示しており、一般的に検証試験であれば、検出力は 75%から 80%以上を設定することが多い。臨床研究では、症例数の設計を行うときに検出力が 80%程度を確保できるように、主要な症例数の見積もりを行う。しかしながら、現実的には診療実績や疾患の特異性から収集可能な症例数にあらかじめ制約を受けていることがあるため、症例数を固定した上で検出力がどの程度を担保できるかを調べることも多い。本研究ではそうした現実的な設定を考慮し、症例数を固定したまま検出力がどのように変動するかを解析した。

本研究では、症例数を固定し、求めたい差の大きさを変化させることで検出力の変化を求めた。求めたい差がごくわずかであり、十分な検出力が得られない場合には、計測値指標を変更するなどの異なる方法の考慮が求められるであろう。

結果の図 1 から、放射線画像が 30 例のときに、AUC の差 0.035 よりも小さい場合には、80%の検出力を確保することは困難である。そこで、読影者を 10 人よりも増やすことで検出力を確保する事が可能である。読影者を 5 人以上に増やす場合、検出力の上がり幅は大きいが、人数を増やすことで検出力の上昇幅に減少がみられる。これは、ROC 実験において読影者を増やしても、ある人数から読影者の増加による効果が減少することを意味している。非心 F 分布は、非心パラメータの λ が大きくなると、確率密度分布が値の大きい方へシフトする。したがって、読影者を増やしていくと非心 F 分布の歪みが大きくなるため同じ AUC の差であっても検出力が大きくなり、やがてプラトーに達する。

Obuchowski らの論文では、相関の関係性について $r_1 \geqq r_2 \geqq r_3$ となることを指摘している[(1)]。もし、異なるモダリティで異なる読影者が設定された場合には $r_1=r_3$ となり、異なるモダリティで症例を読影した場合には、$r_1=r_3=0$ となる。本研究では、$r_1=r_2=r_3=0.1$ とし、それぞれに弱い相関構造を考慮した。これは、複数の読影者が同じモダリティに対して、同一の症例群を読影することを指向している。極端な例を考えると、$r_1=r_3=0$ の場合は、異なる読影者間で、同一の放射線診断機器における相関のみが、被験者間の分散に対する重みとして寄与する。実際には、読影者ごとに読影のくせがあるため、$r_1=0$ は考えにくいが、それぞれの相関が 0 でなければ被験者分散が小さくなり、非心パラメータが大きくなることから、読影者の増加による検出力の増加効果は小さくなる。ただし、r_1 と r_3 が一定という条件のもとで、異なる読影者間で、同一の放射線診断機器における相関である

r_2 が大きい場合には、被験者分散が大きくなり検出力の増加効果を大きくなると考えられる。
今後は、複数のパラメータを調整することで、検出力曲線の変化を描画することを検討している。

6. 結論

本研究で計算した検出力の結果から、AUC の差が 0.035 以下など、小さいと予想される ROC 実験を行う際には、読影者を 10 名以上の協力を得て実施することで、80%以上の検出力を確保できることが明らかになった。また本研究の結果から、10 名以上の読影者の協力を得る場合、検出力の増加は小さくなることが明らかになった

7. 参考文献

(1) Obuchowski NA. Multireader receiver operating characteristic studies: a comparison of study designs. *Academic radiology.* Aug 1995;2(8):709-716.

(2) Obuchowski NA, Beiden SV, Berbaum KS, et al. Multireader, multicase receiver operating characteristic analysis: an empirical comparison of five methods. *Academic radiology.* Sep 2004;11(9):980-995.

(3) Obuchowski NA. New methodological tools for multiple-reader ROC studies. *Radiology.* Apr 2007;243(1):10-12.

(4) Obuchowski NA. Nonparametric analysis of clustered ROC curve data. *Biometrics.* Jun 1997;53(2):567-578.

(5) 白石順二. CAD 研究に役立つ ROC. *医用画像情報学会雑誌.* 2004;21(1):30-38.

(6) 白石順二. ROC 解析における観察者および試料間変動を考慮した統計的有意差検定. *日本放射線技術学会雑誌.* 2007;63(10):1200-1207.

(7) 鈴木信昭, 清野和絵. ROC 解析における Jackknife 法を用いた観察者間変動の解析. *日本放射線技術学会雑誌.* 2010;66(11):1492-1496.

(8) 竹田智, 後藤淳, 丸野達也, 本田拓也, 持留浩輔, 白石順二. Receiver operating characteristics(ROC)解析における信号の選択に関する検討. *日本放射線技術学会雑誌.* 2010;66(11):1467-1473.

(9) Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data*: Springer; 2000.

(10) Oda S, Awai K, Suzuki K, et al. Performance of radiologists in detection of small pulmonary nodules on chest radiographs: effect of rib suppression with a massive-training artificial neural network. *AJR. American journal of roentgenology.* Nov 2009;193(5):W397-402.