ゲノムワイドスクリーニングでの感度評価

リスク解析戦略研究センター，データ科学研究系　多次元データ解析グループ
教授　　松井　茂之

1　はじめに

がんの診断では，がんの形態学的な情報のみならず，がん細胞に特異的な染色体異常やタンパク質・遺伝子の発現異常などの分子レベルの情報も用いられる．近年，がんの進展と病態の背後にある未知のメカニズムを発見するために，ゲノム全体を網羅的に調べる研究（ゲノムワイド研究）が盛んに行われている．そこでは，数千・万の遺伝子の中から，がんの病理診断や予後などの臨床情報と関連のある遺伝子の選抜（遺伝子スクリーニング）が試みられる．

遺伝子スクリーニングのための統計解析として，これまで，多重検定（multiple testing）が主に議論されてきた．数千・万の検定を同時に行うことは，多重検定の極端な例といえるものであり，多重検定の方法論の研究としては大いに興味がもたれる．しかし，がんの研究者は，多少の偽陽性は許しても，重要な関連遺伝子はもれなく選抜したいと望むであろう．これより，検出力の評価が重要になる．しかし，ほとんどの適用例でこれは行われていない．このことは，原理的に偽陽性のコントロールを行う検定の枠組みを採用している以上，当然の成り行きのようにも思える．

一方，疾患の早期発見の分野では，一般健常人にある検査を行って疾患のスクリーニング（disease screening）が行われる．検査法の性能評価では，感度（疾患ありの場合に正しく陽性と判定する確率），特異度（疾患なしの場合に正しく陰性と判定する確率）という二つの能力のバランスが重視される．実際の検査では，あるマーカーの発現量等を測定し，それがある閾値以上であれば陽性と判定する．閾値を変えたときの感度と特異度の変化は，ROC曲線（縦軸：感度，横軸：1－ 特異度）によって捉えられ，感度と特異度のバランスから，適切な閾値が設定される．本稿では，疾患スクリーニングを参考にして，遺伝子スクリーニングのための新しい枠組みの構築を試みる．

2　遺伝子スクリーニングで用いる指標

感度は，「関連ありの遺伝子を選抜する確率」，特異度は，「関連なしの遺伝子を選抜しない確率」と定義する．多重検定の枠組みとの対応では，感度は，関連遺伝子における検出力の平均，１－ 特異度は，個々の検定の有意水準 α に相当する（すべて検定で共通の有意水準を用いることを想定している）．閾値は有意水準 α に相当する．しかし，疾患スクリーニングと遺伝子スクリーニングで異なる点がある．疾患スクリーニングでは，公衆衛生的な観点から，つまり，社会全体での疾患対策やその経済的評価の観点から，（数の上で圧倒的に多い）健常人に関わる特異度を評価する意義を見いだせる．一方，遺伝子スクリーニングでは，特異度は，関連なしの遺伝子に関わるものであり，がんの研究者にとってこれは興味の対象ではない．一方で，がんの研究者は，遺伝子スクリーニングで選抜した遺伝子セットについて，（それが関連遺伝子であると信じて）詳細な生物学的な検討を行うので，遺伝子セットの中に関連なしの遺伝子（偽陽性）が多く

紛れ込んでいることは大変迷惑なことと考えるだろう．そこで，選抜した遺伝子セットにおける関連なしの遺伝子の割合（あるいはその期待値）を偽陽性の指標として用いることが考えられる．これは多重検定の枠組みで提案されている偽発見率（false discovery rate; FDR）に他ならない．通常，ROC曲線には上に凸な非減少な関数を想定できる．さらに，ゲノムワイド研究では関連なしの遺伝子の割合は（未知であるものの）定数と考えてよい．これらの条件のもとで，1 – 特異度と偽発見率は近似的に１対１の関係にあることが示せる．以下では，1 – 特異度の代わりに偽発見率を横軸に用いたROC曲線を用いることを考える．なお，偽発見率の推定に関しては多重検定の枠組みで多くの研究がある．

3　感度の推定

感度の推定では，関連遺伝子全体での効果サイズの分布の推定が必要となる．例えば，二つのクラス（クラス１と２）をもつ臨床変数（例えば，がんの再発ありとなし）と遺伝子 j の発現量の関連づけを考える．効果サイズの指標として，二つのクラス間での発現量の平均の差をクラス内標準偏差で割った

$$\Delta_j = \frac{(\mu_{j,1} - \mu_{j,2})}{\sigma_j}$$

(standardized mean difference; SMD) ($j = 1, \ldots, m$), その推定量として Y_j（例えば，二標本 t-統計量）を用いるとする．全遺伝子から，関連遺伝子を分離し，さらに，サンプリング誤差を考慮した上で，効果サイズの分布を推定するために，次の階層混合モデルを用いる．

$$f(y) = (1 - \pi) f_0(y) + \pi f_1(y)$$

π は混合確率であり，関連遺伝子である確率である．関連なしの遺伝子についての分布 f_0 は平均ゼロの正規分布，関連遺伝子についての分布 f_1 については，次の階層モデルを仮定する．

$$Y_j \mid \Delta_j \sim \mathrm{N}(\Delta_j, \tau_j), \quad \Delta_j \sim G$$

分布 G のノンパラメトリック推定は，smoothing-by-roughening (Laird and Louis (1999))での EM アルゴリムを応用することで得られる．片側 $\Delta_j < 0$ である関連遺伝子のスクリーニングでは，統計量 Y の閾値 y に対して，偽発見率，感度は，

$$\hat{Fdr}(y) = \frac{\hat{\pi}_0 \hat{F}_0(y)}{\hat{F}(y)}, \quad \hat{Se}(y) = \hat{F}_1(y)$$

によって推定できる．ここに F は分布関数である．感度は，全ての関連遺伝子における選抜された遺伝子の割合である．また，研究者にとって特に興味のある，関連が最も強い遺伝子（上位遺伝子）における選抜割合を考えることもできる．これを部分感度（partial sensitivity）とよぶことにする．これは，関連遺伝子の効果サイズの分布 G に基づいて定義できることに注意する．

なお，G の推定値に基づいたサンプルサイズ設計も容易に構成できる．これは，将来のゲノムワイド研究における感度・部分感度のコントロールを可能にするものである．

参考文献

Laird, N. M. and Louis, T. A. (1991). Smoothing the non-parametric estimate of a prior distribution by roughening: a computational study. *Computational Statistics and Data Analysis*, **12**, 27-37.