

## 論文の内容の要旨

論文題目 Regional inference procedure using smoothing techniques for case-control association studies

和訳 ケース・コントロール関連研究における平滑化を用いた疾患感受性遺伝子領域を推測する方法

指導教官 大橋 靖雄 教授

東京大学大学院医学系研究科

平成 13 年 4 月 進学

保健学博士後期課程

健康科学・看護学専攻

氏名 田中 紀子

### はじめに

近年、技術の進歩に伴い、genetic association study でも頻度の高い疾患の疾患感受性遺伝子を見つけるために一度に多くのマーカーを用いられるようになってきている。そのような Whole-genome study を行った場合、通常は疾患との関連を統計的検定を行って判断するため、検定の多重性の問題が生じる。有意水準について今まで様々な議論がなされ、多くの棄却限界値が提案されてきている。検定の多重性を考慮した棄却限界値は必要ではあるが、一つ一つのマーカーが疾患と関連があるか否かを検出するためには、非常に大きなサンプルサイズを要する。そこで、遺伝子データは互いに相関があるため、その相関構造を利用し、領域単位で疾患と遺伝子多型との関連の有無を測定することが提案されている。

関連の有無を判断するためには統計的検定は有用な手段であるが、疫学研究においては関連の強さの程度に興味がある場合が多い。また、推定値の分散や信頼区間幅を検討することで、 $p$  値だけからは得られない情報、つまりマーカー多型頻度の情報も得ることができる。通常ケース・コントロール関連研究においては、関連の強さの指標としてオッズ比や連鎖不平衡値が用いられる。この場合においても、特に whole-genome study においては、個々のマーカーごとの関連の強さよりも、どの領域がどの程度疾患と関連しているか、ということに強く関心があるであろう。しかし、領域単位で疾患と遺伝子多型との関連の強さを測定できる方法は提案されていない。

そこで、本研究では、オッズ比の重みつき平均を計算する平滑化法により、ゲノムのどの領域が最も疾患感受性が強いかを推論できる方法を提案し、さらにその性能評価を行うことを目的とする。また、実際に得られたミトコンドリア DNA の SNPs とアルツハイマー病との関連を調べるために行われたケース・コントロール研究のデータに提案する方法を適用し、アルツハイマー病の発症に関連する遺伝子領域を探索する。

## データ

本研究では late onset アルツハイマー病と様々なリスク要因との関連を調べるためのケース・コントロール研究から得られたケース 613 例、コントロール 408 例のうち、ランダムにサンプリングされて mtDNASNP のデータを得られたケース 154 例、コントロール 129 例を対象とした。MtDNA には 13 の遺伝子領域が存在するが、そのうち最も多くの領域を占めている NADH ヒドロゲナーゼのサブユニットである 7 つの遺伝子領域に存在した 403SNPs を対象とし、シミュレーションのための基データ、および適用例のためのデータとした。

## 提案する方法

### オッズ比

マーカー(あるいは遺伝子)と疾患の関連の強さの指標としてオッズ比を採用した。

### オッズ比の平滑化

周辺の遺伝子のオッズ比の情報を用いたオッズ比の重み付き平均(平滑化オッズ比)を関連の指標として用いることを提案する。今、 $i$  番目のマーカーにおける単純なオッズ比を  $\theta^i$ 、計算される重み  $w^i$  とすると、ある区間  $[i-l, \dots, i+l]$  における平滑化オッズ比は

$$\theta^{i*} = \frac{\sum_{k=i-l}^{i+l} w^k \theta^k}{\sum_{k=i-l}^{i+l} w^k}$$

と計算される。

### 重みの選択

重み  $w^i$  として、マーカー間の相関を考慮できる指標としての連鎖不平衡値の一つである  $D'$ (重み1)、また相関構造が得られない場合も考え、二つ目の重みとしてマーカー多型頻度(重み2)を提案する。

### 信頼区間の計算とバイアス補正

ブートストラップ法(パーセントイル法)により平滑化オッズ比の信頼区間の計算を行った。また、先行研究より、マーカーアリアル頻度が低い場合には一つ目の重みに採用した  $D'$  にバイアスが生じることが知られていることから、平滑化オッズ比においてもブートストラ

ップバイアスで補正した値も計算することとした。

### 欠損値の扱い

genome screening を行う多くの場合、コストとのバランスを考えると、タイピングするマーカーが多くなれば多くなるほど、十分なサンプルサイズを確保することが困難となる。また、事前の多型頻度情報が得られない場合も多い。そのような場合、得られたデータの多くで、関連の強さを測定するための指標を計算できない、つまり推定値が欠損値になってしまう状況もある。そこで、重みの選択だけでなく、欠損値の扱い方によって、提案した方法の性能に差があるかどうかを検討した。欠損値の扱いは以下の3通りを提案した。

方法 1.  $2 \times 2$  分割表のセルのどこかに0頻度が含まれていた場合には、その補正方法として各セルに 0.5 を足して計算する。

方法 2. 欠損があったマーカーの情報はすべて無視する。

方法 3. 欠損があったマーカーにおける粗オッズ比の値に、疾患と関連がない場合の真値である1を代入し重みの計算にのみそのマーカーの情報は用いる。

### シミュレーション

提案した方法の性能を評価するため、アルツハイマー病と mtDNASNPs との関連を調べたケース・コントロール研究から得られた実データをもとに擬似データを発生させ、シミュレーションを行った。シミュレーションの設定は、すべてのマーカーが疾患と関連がなかった場合と、ND6 遺伝子における一つのマーカーが疾患と関連があった場合の二通りを考え、それぞれの設定の下で、提案した平滑化オッズ比と、単純に個々のマーカーごとに重みをつけずに推定した粗オッズ比(既存の方法)について、推定値のバイアスを比較した。欠損値の扱いについては、方法1~3を用い、相互に比較を行った。また、ブートストラップ法を用いて計算した信頼区間の coverage rate、区間幅の長さについても検討を行った。

バイアスの大きさと coverage rate は重みの違いよりも欠損値の扱い方の違いにより差がみられ、すべての方法で正のバイアスが生じていたが、相対的には方法3が最もよい性能を示した。これに対し、区間幅の長さは重みによる差がみられ、どの欠損値の補完方法においても、重み1の方が平均的にせまい信頼区間幅であった。また、信頼区間に1を含まないことを疾患と関連があることの基準とした場合の検出力を計算した結果、方法1と方法3で既存の方法より高い検出力を得られた。ただし、真の相対リスクを4に、どの方法・重みを選択しても、平滑化オッズ比の推定値は負の方向に大きなバイアスが生じていた。

平滑化の重みの選択と欠損値の扱いを考慮した結果を総合的に判断すると、特に多型頻度が低く、マーカー間の相関が極端に高くないデータに対しては、方法3で重み

1を採用した場合が精度が高く、保守的ではあるが、比較的検出力あり、最も性能がよいことが示唆された。

### 提案した方法の実データへの適用

提案した方法をアルツハイマー病のデータに適用し、推定値のプロットをグラフ化し、どの領域で多型の頻度が高く、どの領域で疾患と関連がある可能性があることを視覚的に判断できることが示せた。粗オッズ比を計算し、その信頼区間が1を含まないマーカーを疾患との関連があるとした場合 11 のマーカーで関連がみられたが、シミュレーションの結果最も性能がよかった方法3で重み1を採用した場合にはすべてのマーカーで関連がみられなかった。これらより、ミトコンドリアの NADH ヒドロゲナーゼのサブユニットである 7 つの遺伝子領域においては、アルツハイマー病と関連の強い遺伝子はないと考えられ、比較的弱い(オッズ比にして2程度)の領域も存在する可能性が低いことが示唆された。

### 考察

本研究で提案した方法は非常に単純な方法であり、拡張性が高く、今回適用したようなデータではなく、アリル頻度が全体的に高い場合や、よりマーカー間の相関が高い場合にも適用可能である。しかし、シミュレーションの結果から、提案した方法は相関構造の変化に敏感である可能性も示唆されている。適用したデータに対しては方法3で重み1を採用した場合が最も性能がよいことが示唆されたが、他の方法も適用するデータの構造や、信頼区間の計算方法を改善することにより、性能がよくなることも考えられる。今後他のデータにも適用し、更なる検討が必要と考えられる。

### 結論

オッズ比の重みつき平均を計算する平滑化法により、ゲノムのどの領域が最も疾患感受性が強いかを推論できる方法を提案した。今回用いたデータのような相関構造をもつ領域に関しては、既存の方法に比べ提案する方法の性能は高く、偽陽性のマーカーを減らし、疾患感受性変異の検出能力においても優れていることが示唆された。