科学研究費助成事業　　研究成果報告書

平成 ２６ 年　 ６ 月　 ３ 日現在

機関番号：１３１０１
研究種目：挑戦的萌芽研究
研究期間：2012～2013
課題番号：２４６５０１４９

研究課題名（和文）遺伝子発現データに基づく予後予測モデル構築の統計理論の開発と実データによる検証

研究課題名（英文）Statistical theory and the evaluation of Cox's proportional hazards model regularized by various penalties for survival DNA microarray expression data

研究代表者
赤澤　宏平（Akazawa, Kohei）

新潟大学・医歯学総合病院・教授

研究者番号：１０１７５７７１

交付決定額（研究期間全体）：（直接経費）　　2,400,000円、（間接経費）　　720,000円

研究成果の概要（和文）：　本研究では、疾患関連遺伝子の生存時間に与える影響を評価する際に、予後規定因子を確率的に正しく検出する罰則付き回帰モデルを開発しその性能を評価した。
　シミュレーションの結果から次のことがわかった。（１）候補因子の個数が症例数より多い時、罰則付きＣｏｘモデルでの検出力は逐次変数増加法に比べて高い傾向にある。（２）この傾向は、罰則項にL1-ノルム、L2-ノルムを用いた際に顕著になる。（３）実際のＤＮＡマイクロアレイデータに本手法を用いたところ、疾患感受性遺伝子の検出力を正しく定量的に推定できた。今回開発した罰則付きＣｏｘモデルは、遺伝子解析データの生存予後解析を行う際に有用なツールとなりうる。

研究成果の概要（英文）：In this study, we created programs using the R language to calculate statistical properties of Cox's proportional hazards model regularized by various penalties, including the statistical power based on the prognostic index, and conducted simulation experiments under various conditions of DNA microarray expression data with survival time. The results showed that when the number of candidate genes is larger than the sample size, the power of a validation set for penalized methods is greater than for stepwise methods in many cases. This tendency is most remarkable for the penalized methods including both the L1-norm and the L2-norm. Furthermore, we verified our programs to actual microarray gene expression data with survival time data to confirm their validity. Our simulation programs for Cox's proportional hazards model regularized by various penalties are very useful for planning DNA microarray studies or for evaluating the results of such studies.

研究分野：総合領域

科研費の分科・細目：情報学・統計科学

キーワード：比例ハザードモデル　罰則化　DNAマイクロアレイデータ　生存時間解析

１．研究開始当初の背景

（１）DNA マイクロアレイ解析によって得られる腫瘍組織の遺伝子発現プロファイルには、癌細胞で生じるシグナル経路の変異が反映されている。したがって、そのデータと臨床所見の統合によりがん患者の予後予測が可能となる。実際に、早期浸潤性乳癌の患者の手術時摘出組織について、21 個の遺伝子の発現データと臨床所見（腫瘍径、臨床進行期、病理学的所見など）をモデル化することにより、10 年間の患者の予後予測が実用化されている。数万個の遺伝子の中から、予後予測に最適な遺伝子を選択する手法、ならびに、これらの遺伝子を用いて予後予測スコア等を算出するモデルを構築することは、医学統計学に課せられた大きな課題といえる。

（２）遺伝子発現データと臨床所見との統合による癌患者の予後予測は、以下の手順で行なわれることが多い：① 研究対象のトレーニング群とバリデーション群への無作為割付、② トレーニング群を用いた遺伝子の同定、③②で同定された遺伝子情報と臨床所見を組み合わせた最適な予後予測モデル（判別器）の構築、④構築された予測モデルのバリデーション群への適用と判別能力の評価。この過程の具体的なプロセスを調べてみると、既存の代表的な研究でも用いられる研究デザインと統計手法が統一されていないことがわかった。

（３）特に、トレーニング群における予後規定遺伝子の同定方法はそれぞれの臨床論文で独自の方法が用いられており、複数の研究結果の比較も行うことができない状況にある。権威ある論文誌の Nature Genetics, Lancet, NEJM などでも、各研究グループが独自の方法で遺伝子の探索を行っており、再現性のある確証的な結果が得られているとは言い難い。このことは Simon R らも 2003 年、2007 年の JNCI で指摘している。

（４）DNA マイクロアレイを用いた遺伝子発現データと臨床所見とを統合して、がん患者の生存・再発予後を予測する統計学的手法は、国際的に見て未だ確立されているとは言えない。

２．研究の目的

（１）本研究では、生存予後に有意な影響を与える疾患関連遺伝子を検出するための既存の手法を再評価して、数万個の候補遺伝子からの予後規定遺伝子の特定、および、予後予測モデルの構築を行うための統計学的手法を開発する。

（２）本研究実施に際しては、実際の遺伝子発現データに基づく統計学的問題点の発掘が行なわれなければならない。また、生存時間解析の理論の展開と大規模なシミュレーション研究が必要となる。本研究では、生存時間解析の研究者とがん診療・DNA マイクロアレイの専門家が、理論・手法の開発、シミュレーションによる検証および臨床への応用を共同で検討する。

３．研究の方法

（１）本研究の主たる研究の進め方は次のとおりである。

①マイクロアレイによる遺伝子発現プロファイル解析の問題点の発掘

②実データの解析による問題点の検出と統計学的知見の整理

③問題点解決のための理論的アプローチならびにシミュレーションによる検証

医学統計学の専門家とがん診療・遺伝子発現プロファイルの専門家が、同一施設内にて、上述のプロセスを共同ですすめた。これ以外に、統計解析用ソフトウエア SAS や R 言語に精通した共同研究者が統計処理やシミュレーションプログラムの開発を行なった。本研究の大きな特長は、実際に収集される卵巣癌患者の遺伝子発現データを利用しながら統計理論の展開、コンピュータシミュレーションによる検証および臨床への応用を同時に実施できる点にある。研究方法の詳細は以下のとおりである。

a）遺伝子解析データのうち、本研究ではマイクロアレイ解析データを用いることにする。特に、すでに疾患感受性遺伝子の検出が行われ国際学術雑誌に掲載されているがん患者のマイクロアレイデータを用いる。

b）このデータセットの特性を統計学的に分析し、それらの特性をパラメータ化する。

c）b）の種々の条件下でモンテカルロシミュレーション用のデータセットを作成する。データセットにはトレーニングデータセットとバリデーションデータセットの２種類がある。

d）b）の種々の条件の下で、指数モデルを用いて比例ハザードモデルに従う生存時間乱数を生成する。マイクロアレイデータと生存時間乱数を結合してトレーニングデータセットとする.

e）トレーニングセットを用いて、罰則化項を有しない通常の最小二乗法（LSM)、罰則化項を有する ridge ペナルティ、elastic net ペナルティ、lasso ペナルティによる比例ハザードモデルの推定値を算出する.

f)トレーニングセットと同じ要領で別の生存時間乱数を生成し、マイクロアレイデータと結合したデータをバリデーションデータセットとする.

g）トレーニングデータセットから得られた情報、すなわちトレーニングデータセットで選択された遺伝子とその回帰係数とを用いて、バリデーションデータセットの各症例の予後指数（当該遺伝子の発現値と回帰係数との積の和）を算出する.

h）バリデーションデータセットを予後指数の中央値により高リスク群と低リスク群の 2 群に分け、2 群の生存率曲線を log-rank 検定で検定する.

i）以上の手続きを 2000 回行い、検出力（P 値が 0.05 未満になる割合）を算出する.

j）各種手法の妥当性を評価するために、バリデーション率（トレーニングデータセットとバリデーションデータセットの両セット間において、高リスク群と低リスク群との 2 群間での有意差検定の判定結果が一致する割合）、感度（選択された遺伝子中の真の遺伝子数／真の遺伝子数）ならびに陽性反応適中度（選択された遺伝子中の真の遺伝子数／選択された全遺伝子数）を算出する.

４．研究成果

（１）　検出力はいずれの手法においても症例数の増加と共に増加し、候補遺伝子数の増加と共に減少した．トレーニングデータセットとバリデーションデータセットとの比較では、後者の検出力は前者に比べて常に低くなっていた．データセット別では、トレーニングセットでは LSM 法の方が elastic net 法より検出力が高くなっていたが、バリデーションセットでは、elastic net 法が LSM 法より高い検出力を示した。
感度は、いずれの手法においても症例数の増加と共に増加したが、LSM 法は elastic net 法よりも常に高い値となっていた．陽性反応適中度は、症例数の増加と共に LSM 法は減少したが、elastic net 法は増加し、elastic net 法は LSM 法よりも常に高くなっていた．バリデーション率はいずれの手法においても症例数の増加と共に増加したが、elastic net 法は LSM 法よりも常に高い値を示した(図1).

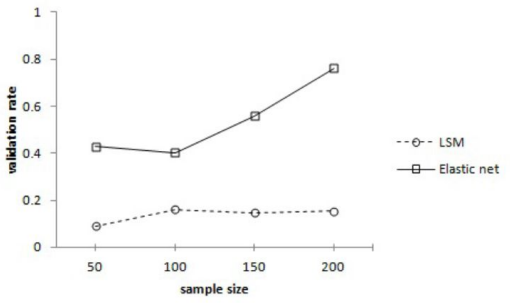

図１．ＬＳM法と Elastic net 罰則化法におけるバリデーション率の比較

（２）上述の結果以外に、罰則化項を有しない通常の最小二乗法（LSM)、罰則化項を有する ridge ペナルティ、elastic net ペナルティ、lasso ペナルティによる比例ハザードモデルによる解析の比較を、選択された遺伝子数、選択された遺伝子の回帰係数の推定値、選択された遺伝子の回帰係数の標準偏差、陽性反応適中度および検出力について、理論的考察と詳細なシミュレーションを行い、それらの結果を国際的な統計学雑誌に投稿している。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 2 件）
①Wen Y, Miyashita A, Kitamura N, Tsukie T, Saito Y, Hatsuta H, Murayama S, Kakita A, Takahashi H, Akatsu H, Yamamoto T, Kosaka K, Yamaguchi H, Akazawa K, Ihara Y, Kuwano R.
SORL1 is genetically associated with neuropathologically characterized late-onset Alzheimer's disease.
J Alzheimers Dis. 2013; 35(2): 387-394.
査読有

②Yoshihara K, Tsunoda T, Shigemizu D, Fujiwara H, Hatae M, Fujiwara H, Masuzaki H, Katabuchi H, KawakamiY, Okamoto A, Nogawa T, Matsumura N, Udagawa Y, Saito T, Itamochi H, Takano M, Miyagi E, Sudo T, Ushijima K, Iwase H, Seki H, Terao Y, Enomoto T, Mikami M, Akazawa K, Tsuda H, Moriya T, Tajima A, Inoue I, and Kenichi Tanaka for The Japanese Serous Ovarian Cancer Study Group.
High-Risk Ovarian Cancer Based on 126-Gene Expression Signature Is Uniquely Characterized by Downregulation of Antigen Presentation Pathway.
Clin Cancer Res 2012; 18: 1374-1385.
査読有

〔学会発表〕（計 1 件）
①北村信隆、赤澤宏平：疾患感受性遺伝子同定のためのマイクロアレイデータ解析法に関する研究―elastic net 罰則化項により最適化された Cox の比例ハザードモデルの特性について－
第 32 回日本医療情報学連合大会　2012 年 11 月　新潟・朱鷺メッセ

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）

名称：
発明者：

権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
赤澤　宏平（AKAZAWA, Kohei）
新潟大学・医歯学総合病院・教授
研究者番号：10175771