様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　２０　日現在

機関番号：　８２６０６
研究種目：　若手研究(B)
研究期間：　2015～2017
課題番号：　１５Ｋ１５９５６

研究課題名（和文）予測モデル構築のための研究デザインと予測性能の評価方法

研究課題名（英文）Statistical methods to develop and evaluate risk prediction models

研究代表者
口羽　文（Kuchiba, Aya）

国立研究開発法人国立がん研究センター・研究支援センター・室長

研究者番号：４０５１０６９９

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：がんの発症原因が次々と発見されている．リスクとなる遺伝的要因や生活習慣などからがんの発症リスクを予測できれば，リスクに基づく予防法の選択や予防介入の実現に重要な役割を果たすと考えられる．
本研究では，SNPなどの分子マーカーを用いる研究でよく用いられる研究デザインに対する予測性能指標の推定方法を検討し，実データへ応用した．また，関連して，サブタイプによるリスク因子の影響の違いを評価するための方法論とその応用研究を進め，学会・論文発表を行った．さらに，3カテゴリ以上のアウトカムの予測に対するconcordanceの指標を新たに定義し，既存の予測評価指標との違いをシミュレーション研究により評価した．

研究成果の概要（英文）：Recent cancer epidemiology has provided much evidence for cancer risk factors, including environmental, lifestyle and genetic factors. Prediction model to provide the risk of cancer based on the established risk factors, would contribute to developing risk-based prevention or health management strategies.
In the research project, the methods to estimate prediction ability of the prediction models has been examined and applied to real epidemiology studies, in particular for data from often-used study designs with biomarker data. The statistical methods for assessing etiologic heterogeneity of a cancer have been developed and applied to the cohort study data as a relevant topic. In addition, new index to evaluate prediction ability for outcome with more than two categories has been proposed and the characteristics of this index have been examined by simulation studies.

研究分野：　生物統計学/疫学

キーワード：　予測モデル　がん罹患　Calibration　Discrimination　疫学研究

１．研究開始当初の背景

がんの発症原因となる要因が次々と発見されている．近年の common disease に対する遺伝的リスク要因を探索する研究の発展は目覚ましく，生活習慣や環境要因に加え，多くの遺伝的多型が同定されている．また一方では，各部位のがんは，腫瘍組織の分子マーカーデータに基づいてより詳細なサブタイプに分類されるようになった．リスク因子の探索においても，分子サブタイプによってリスク因子やその効果が異なることが示唆されている．リスク因子とその効果の大きさの確立すること，またそれらの情報からがんの発症リスクを予測することは，リスクに基づく予防法の選択や予防介入の実現のために重要の課題の一つであると考えられる．

しかし，予測性能を最適化するための研究戦略の検討は，関連の評価を目的としたものに比べまだ少なく，研究方法論の発展，方法論的枠組みの提案が望まれている．

２．研究の目的

発症予測の観点から，研究デザイン・統計学的手法の検討を行うことを目的とする．

さらに，検討・開発した手法は，実際の研究データへ応用し，疫学的な研究課題へも貢献することを目指す．

３．研究の方法

疫学研究データを用いた発症予測モデルの構築や評価にかかわる統計的課題に取り組む．

統計手法や指標の評価は，理論的な導出とシミュレーション研究により行う．さらに，これらを実際の疫学研究データへ適応し，主要部位のがんに対する予測モデルの構築・評価を行う．

４．研究成果

(1)予測モデルの性能評価：

①予測モデルの良さは calibration と discrimination の２つの面から評価される．Calibration は，モデルから予測される疾患発症確率（予測確率）と実際のリスクとの一致の程度を表す．Discrimination は，予測モデルが，将来疾患を発症する人としない人をどの程度区別できるかについての指標である．よく用いられる ROC 曲線の曲線下面積（AUC：Area Under the Curve）は，discrimination を評価するための代表的な統計量である．

SNP などの分子マーカーを用いる研究では，ケース・コホートデザインやネスティッドケース・コントロールデザインのように，コホート全体集団から一部の集団をサンプリングするデザインが用いられることが多い．そこで，このようなサンプリングデザインにおける calibration, discrimination 指標の推定方法に関する検討を行った．

さらに，これらの指標を大腸がん罹患予測モデルの評価に適応した．この研究の目的は，生活習慣による既存の発症予測モデルに遺伝的なリスク因子を加えることで予測性能が向上するかどうかを評価することであった．遺伝的なリスク因子である SNP は，ネスティッドケース・コントロールデザインで評価されている．実際に，本研究データを用いて予測モデルの構築と評価を行い，SNP を加えることで予測性能が向上することが示唆された．

②多くの予測性能指標は，予測したいアウトカムが「発症する/しない」のように 2 カテゴリである場合を対象としている．一方で，今後，興味のあるアウトカムが 3 カテゴリ以上（例えば，発症なし，ER+乳がん，ER-乳がん）となる例は増えていくと考えられるが，2 カテゴリに対する予測と比べると，その予測性能の評価は複雑になる．

例えば，図 1 の上のグラフは，アウトカムが 2 カテゴリの場合の（Y=1 or 2），各アウトカム群における予測確率（p_1：アウトカム Y=1 に対する予測確率）の分布とする．この場合，赤と緑の分布が離れているほど，予測モデルの性能は良いといえる．一方で，図 1 の下の図は，アウトカムが 3 カテゴリの場合の同様のグラフであり（Y=1, 2, or 3），各アウトカム群における予測確率の分布を 3 次元で示している（p_1：アウトカム Y=1 に対する予測確率, p_2：アウトカム Y=2 に対する予測確率）．2 カテゴリのアウトカムの場合と同様，この赤，青，緑の 3 つの分布が離れているほど，予測モデルの性能は良いといえるが，「離れている」の定義は 2 カテゴリの問題と比べると単純ではない．また，予測モデルの適応場面やアウトカムカテゴリ間の関係に応じて，複数の指標による多面的な評価が必要になると考えられる．

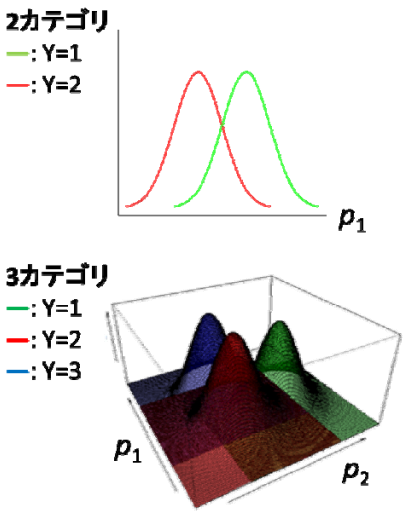

図 1．2 カテゴリアウトカムと 3 カテゴリアウトカムの予測確率の分布（仮想例）

AUC の考え方は，3 カテゴリ以上のアウトカムに対しても拡張されている．本研究では，3 カテゴリ以上のアウトカムの予測に対して，新たに concordance の指標を考案した．予測モデルからそれぞれのアウトカムカテゴリに対する予測確率が得られるものとする．実際に観察されたアウトカムが異なる対象者からなるセットを考え，その対象者間での各アウトカムカテゴリに対する予測確率の大きさの関係を用いて定義した．提案指標は，アウトカムカテゴリの間に順序などの特別な関係性を仮定することなく得られる指標である．

シミュレーション研究により，提案指標の特徴，既存の指標との振る舞いの違いを検討し国際学会にて発表した．今回新たに提案している指標については，今後さらに詳細な検討と拡張，実データへの応用研究を続ける予定である．

(2) がんサブタイプに対するリスク因子の評価：

関連する研究として，がんの分子サブタイプ間で，リスク因子が異なるかどうかを評価するための統計手法の研究も進めている．競合リスク解析の枠組みを用い，曝露因子 (X) とサブタイプとの関連をそれぞれモデル化した (図 2)．その各関連 (β) より，例えば，i 番目のサブタイプと j 番目のサブタイプとの曝露効果の違いは $\beta_i - \beta_j$ (heterogeneity parameter) と定義することができる．疫学研究でよく用いられるいくつかの研究デザインの下で，heterogeneity parameter を推定，検定する方法を提案した．

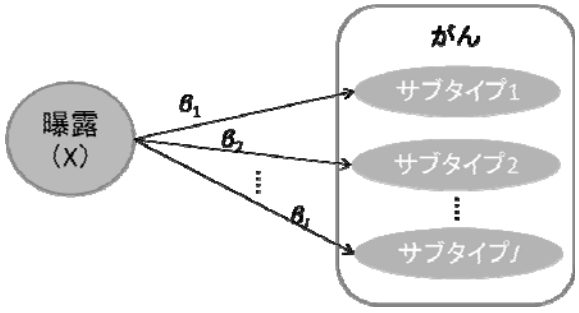

図 2．曝露因子と各サブタイプとの関連

実際に，このアプローチをコホート研究データに適応し，同時性の大腸がんと同時性ではない (通常の) 大腸がんとで，喫煙の影響が異なること示した．

また，この枠組みを，分子サブタイプが，複数の分子マーカーから定義される状況へ拡張した．例えば，遺伝子変異あり・なしのように 2 値の分子マーカー3 つの組み合わせで分子サブタイプが定義されるとすると，可能性としては 2^3=8 通りのサブタイプが存在することになる．分子マーカーが増えるほど，潜在的な分子サブタイプの数は膨大となり，解析モデルに工夫が必要となる．このような状況において，どの分子マーカー変化が，曝露因子の効果の違いを反映しているのかを検討するために，前述した枠組みを拡張し，他の分子マーカーの影響を制御した下で，各分子マーカー変化が，あるリスク因子による etiologic heterogeneity と関連しているかどうかを検討するための方法を提案した．

さらに，この heterogeneity parameter は，がんケースのみ (case only デザイン) からも推定することができる．Case only デザインは，適切なコントロール集団の設定が難しい場合や，新しい分子マーカーの測定などで追加のコストが必要になる場合には，特に有用なデザインである．Case only デザインでも，heterogeneity parameter に対する妥当な推論が可能であることを理論的に導出するとともに，ケース・コントロール研究と比較して，推定効率がどのように異なるかを数値実験，シミュレーション研究により評価した．

(3) 寄与割合の分解

疫学研究におけるリスク因子の評価は，相対リスクの大きさだけでなく，そのリスク因子の集団に対するインパクトとして，寄与割合を評価することも重要である．寄与割合とは，そのリスクを取り除いたら減らすことができるであろう罹患者の割合である．

近年，遺伝的リスク要因が次々と同定されているが，遺伝的な要因自体を取り除くことはほぼ不可能である．しかし，この遺伝的なリスク要因と交互作用する modifiable な曝露要因 (生活習慣など) がある場合には，この曝露要因を取り除くことで，遺伝的要因の寄与の一部も除くことができると考えられる．

2 つのリスク因子とそれらの交互作用の 3 つの要素を考える．潜在アウトカムという考え方の枠組みを用いて，2 つのリスク因子全体の寄与を，各リスク因子の主効果部分とそれらの交互作用部分へと分解する方法を提案した．

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　6　件）

① Taguri M, Kuchiba A. Decomposition of the population attributable fraction for two exposures. Annals of Epidemiology. 2018; 28(5):331-334 (doi:10.1016/j.annepidem.2018.02.012)

② Budhathoki S, Hidaka A, Yamaji T, Sawada N, Tanaka-Mizuno S, Kuchiba A, Charvat H, Goto A, Kojima S, Sudo N, Shimazu T, Sasazuki S, Inoue M, Tsugane S, Iwasaki M for the Japan Public Health Center-based Prospective Study Group. Plasma 25-hydroxyvitamin D concentration and subsequent risk of total and site-specific cancers in a Japanese population: A large case-cohort study within the Japan Public Health Center-based Prospective Cohort. BMJ. 2018; 7;360:k671. (doi:10.1136/bmj.k671)

③ Iwasaki M, Tanaka-Mizuno S, Kuchiba A, Yamaji T, Sawada N, Goto A, Shimazu T, Sasazuki S, and Tsugane S; for the JPHC Study Group. Inclusion of a Genetic Risk Score into a Validated Risk Prediction Model for Colorectal Cancer in Japanese Men Improves Performance. Cancer Prevention Research. 2017; 10(9):535-541. (doi:10.1158/1940-6207.CAPR-17-0141)

④ Drew DA, Nishihara R, Lochhead P, Kuchiba A, Qian ZR, Mima K, Nosho K, Wu K, Wang M, Giovannucci E, Fuchs CS, Chan AT, Ogino S. A Prospective Study of Smoking and Risk of Synchronous Colorectal Cancers. American Journal of Gastroenterology. 2017; 123(3) 493-501. (doi:10.1038/ajg.2016.589)

⑤ Wang M, Spiegelman D, Kuchiba A, Lochhead P, Kim S, Chan AT, Poole EM, Tamimi R, Tworoger SS, Giovannucci E, Rosner B, Ogino S. Statistical Methods for Studying Disease Subtype Heterogeneity. Statistics in Medicine. 2016; 35(5):782-800. (doi: 10.1002/sim.6793)

⑥ Wang M, Kuchiba A, Ogino S. A Meta-Regression Method for Studying Etiologic Heterogeneity across Disease Subtypes Classified by Multiple Biomarkers. American Journal of Epidemiology. 2015; 182: 263-70. (doi: 10.1093/aje/kwv040)

〔学会発表〕（計　5　件）

① Kuchiba A, Sakamaki K. Discrimination Index for Multi-Category Outcome. ENAR 2018 Spring Meeting, in Atlanta, Georgia, USA, March 25-28, 2018.

② Wang M, Kuchiba A, Gao R. Efficiency Consideration of the Etiological Heterogeneity Evaluation in Case-Case and Case-Control Studies. ENAR Spring Meeting, Washington, D.C., March 12-15, 2017

③ Drew DA, Nishihara R, Lochhead P, Kuchiba A, Qian ZR, Mima K, Nosho K, Wu K, Wang M, Spiegelman D, Giovannucci EL, Fuchs CS, Ogino S, Chan AT. A prospective study of smoking habit and risk of synchronous colorectal cancers. AACR Annual Meeting 2016, New Orleans, Louisiana, USA, April 16 - 20, 2016. (Abstract Number: 4348)

④ Wang M, Spiegelman D, Kuchiba A, Rosner B, Ogino S. Statistical Methods for Studying Disease Etiologic Heterogeneity. 2016 Epidemiology Congress of the Americas, Miami, Florida, June 21-24, 2016

⑤ Taguri M, Kuchiba A. Decomposition of the population attributable fraction for two exposures. The East Asia Regional Biometrics Conference 2015, Fukuoka, Japan, December 20-22, 2015

〔図書〕（計　1　件）

① Kuchiba A. Evaluation of Cancer Risk in Epidemiologic Studies with Genetic and Molecular Data (p.297-313). In Matsui, S. and Crowley, J. (Eds.), Frontiers of Biostatistical Methods and Applications in Clinical Oncology. Springer Singapore. 2017 (Book Chapter)

６．研究組織
(1)研究代表者
口羽　文　(KUCHIBA, Aya)
国立がん研究センター・研究支援センター生物統計部・室長
研究者番号：40510699