

科研費シンポジウム

「バイオ統計学の挑戦と貢献」報告書

開催責任者：米本 孝二（久留米大学バイオ統計センター）

日時：2015年2月2日（月）～2月4日（水）

場所：九州大学医学部百年講堂 中ホール3（TEL 092-642-6257）

本シンポジウムは

科学研究費・基盤研究(A)「非対称・非線形統計理論と経済・生体科学への応用」

研究代表者：谷口正信（早稲田大学基幹理工学部）

の助成により開催された。

プログラム

2月2日（月）

13:00-13:05 Opening Remarks

セッション1（座長 西川 正子（東京慈恵会医科大学））

13:05-13:50 岡田 栄作（浜松医科大学 健康社会医学講座）

「Evidence-Based Practice を支えるためのバイオ統計学の挑戦と貢献： - 根拠に基づいた政策を実施するためには何が必要か： J A G E S プロジェクトでの研究事例から - 」

13:50-14:35 大谷 敬子（広島大学原爆放射線医科学研究所），富田 哲治（県立広島大学経営情報学部），佐藤 健一（広島大学原爆放射線医科学研究所），原憲行（広島大学原爆放射線医科学研究所），川上 秀史（広島大学原爆放射線医科学研究所），瀧原 義宏（広島大学原爆放射線医科学研究所），星 正治（広島大学名誉教授），大瀧 慈（広島大学原爆放射線医科学研究所），

「残留放射線による健康影響とその主要な曝露源：広島入市被爆者のコホートデータ（1970-2010）を用いた固形がん超過死亡危険度について」

特別講演1（座長 米本 孝二（久留米大学バイオ統計センター））

14:50-15:40 有馬 久富（滋賀医科大学アジア疫学研究センター、シドニー大学ジョージ国際保健研究所）

「シドニー大学ジョージ国際保健研究所における大規模臨床試験」

セッション2 (座長 江口 真透 (統計数理研究所))

16:00-16:45 大山 哲司 (大分大学医学部)

「処置前後に観察された連続 Bounded Outcome Scores に対する
2 変量反応モデル」

2月3日 (火)

セッション3 (座長 赤澤 宏平 (新潟大学医歯学総合病院))

9:30-10:15 平川 晃弘 (名古屋大学医学部附属病院)

「がん第Ⅰ相試験における用量探索法の最近の展開」

10:15-11:00 野間 久史 (統計数理研究所), 五所 正彦 (筑波大学医学医療系)

「Multiple Imputation 法のモデル選択規準」

11:00-11:45 江口 真透 (統計数理研究所), 小森 理 (統計数理研究所)

「2 群判別のための多重マーカーの一般化 t-統計量について: セミパラ
メトリックスとラッソ」

ランチタイム (11:45-13:30)

セッション4 (座長 平川 晃弘 (名古屋大学医学部附属病院))

13:30-14:15 吉田 寿子 (久留米大学バイオ統計学 博士4年, 九州大学大学院 包括
的腎不全治療学), 川口 淳 (京都大学大学院医学研究科), 山下典
生 (岩手医科大学医歯薬総合研究所), 鶴屋 和彦 (九州大学大学院
包括的腎不全治療学)

「脳画像データ解析のための RBF-sPLS の適用」

14:15-15:00 井元 清哉 (東京大学)

「次世代シーケンスデータによる体細胞変異・高多様性領域の解析」

特別講演2 (座長 米本 孝二 (久留米大学バイオ統計センター))

15:30-16:30 二宮 利治 (九州大学大学院医学研究院附属総合コホートセンター)

「わが国の生活習慣病とその合併症の現状と課題: 久山町研究」

16:30-17:30 久保 充明 (理化学研究所統合生命医科学研究センター)

「オーダーメイド医療研究の最前線」

18:00-20:00 懇親会 (百年講堂内カフェテリアにて)

2月4日（水）

セッション5（座長 矢原 耕史（久留米大学バイオ統計センター））

9:30-10:15 佐藤 俊太郎（長崎大学病院 臨床研究センター）

「地方大学における臨床研究支援部門の取り組み ～統計家の視点から～」

10:15-11:00 西川 正子（東京慈恵会医科大学），森川 敏彦（元久留米大学）

「IntCox は無増悪期間解析の Remedy になり得るか」

11:00-11:45 植木 優夫（東北メディカル・メガバンク機構）

「ゲノムワイドな遺伝子 x 遺伝子間相互作用の探索」

11:45-11:50 Closing Remarks

Evidence-Based Practice を支えるためのバイオ統計学の挑戦と貢献：

- 根拠に基づいた政策を実施するためには何が必要か：JAGES プロジェクトの研究事例から-

浜松医科大学 健康社会医学講座

岡田 栄作

【発表要旨】

バイオ統計学は「ライフサイエンスの研究全般を網羅する数理学的研究のこと」と定義されている（久留米大学バイオ統計センターHP）が、我が国においてバイオ統計学者が従事する研究は、EBM（Evidence-Based medicine）を支えるための研究が大半であり、エビデンスに基づく政策・実践（EBP：Evidence-Based Policy & Practice）については議論される機会が少ない。

しかし、近年は高齢化社会への対応を迫られ、社会保障に使える資源は限られており、効果のない施策や技術に資源配分するのではなく、施策にも、より効果や効率が求められる時代になってきた。EBP を実践するためには、エビデンスを得るためのデータベースなどの基盤整備やプログラム評価、政策評価研究などの蓄積が必要である。

これらの課題に取り組むために厚生労働科学研究費補助金などで組織された JAGES（Japan Gerontological Evaluation Study, 日本老年学的評価研究）プロジェクトでは自治体と共同研究を実施し、研究参加自治体の介護予防政策の立案・実践を支援してきた。2015 年 1 月現在参加自治体は全国 30 市町村である。

本研究プロジェクトでは、要介護認定を受けていない 10 万人超の高齢者のデータベースを構築し、指標群を作成して、保険者に結果をフィードバックするベンチマーク・システムを開発した。さらに縦断研究で解明されたリスク要因や保護的要因、応用研究としての介入研究などに基づく指標群の妥当性の検証などを行ってきた。

実際は、地方自治体が、自らの重点課題やその改善方法についてエビデンスに基づいた政策を立案・実行することは決して容易ではない。その原因として挙げられるのが、そうした取り組みを支援するツールが皆無だったことにある。

JAGES では、WHO 神戸センターが開発した Urban HEART（Urban Health Equity Assessment and Response Tool）を参考にして、全国 25 保険者 30 市町村のデータをもとに、22 のコア指標と 18 の推奨指標からなる介護予防のための見える化システム「JAGES HEART」を開発した。このシステムは、介護保険者（市町村および広域連合）が自らの現状を把握して政策を立案・実行し、改善状況をモニタリングしてその効果を検証し、さらなる改善につなげるというサイクルを回すうえで有用なツールとなりえる。

保険者による JAGES HEART の活用プロセスは、①保険者が取り組むべき重点課題の設定、②保険者内における重点対象地域の設定、③介入施策の立案とプログラムの実施、④政策による効果の評価の 4 段階からなる。この 4 段階のプロセスにおけるこの 4 段階のプロセスを総称し、「地域診断システム」と呼んでいる。

JAGES HEART の意義は、客観的で比較可能な数値による「見える化」を通して、保険者職員が地域診断を行い課題に取り組む過程を、評価まで含めて総合的に支援することにある。市町村の主体性や知識を重視しながら、エビデンスに基づく介護予防の取り組みを支援する JAGES HEART に期待されるところは大きい。

本発表では、介護保険者である自治体が、自らの課題の発見・設定と改善取組を支援する具体的な過程を経て、政策策定に至る事例を研究成果を交えて紹介する。また、こういった EBP を実践するためにバイオ統計家は何をサポートし、どのようなスキルが必要になるのか。根拠に基づいた政策・実践を行うためにバイオ統計家が自治体でも専門職員として雇用される社会的要請・期待に至る現状を説明する。

残留放射線による健康影響とその主要な曝露源：広島入市被爆者のコホートデータ
(1970-2010)を用いた固形がん超過死亡危険度について

Main source and effects of residual radiation: excess mortality risk from solid
cancers among Hiroshima A-bomb early entrants cohort data, 1970-2010

大谷敬子¹, 大瀧 慈¹, 富田哲治², 原 憲行¹, 川上秀史¹,
瀧原 義宏¹, 星 正治³, 佐藤健一¹

¹広島大学原医研, ²県立広島大学, ³広島大学名誉教授

【背景】 原爆放射線による被爆は、以下の2つに分けて考えることができる。爆発と同時に放射されたγ線とそれに続いて放射される中性子線による直接的な被爆と、残留放射線による間接的な被爆である。残留放射線は主として中性子線が土壌や建物が中性子により放射化したことによる誘導放射線と核分裂生成物やウラン・プルトニウムなどの原爆材料を含んだ降下物質によるものがある。広島・長崎の原爆被爆者について、初期被爆線量の健康影響、とくに発がん和被爆線量の関係は、これまで放射線影響研究所などが中心となり詳しく調べられてきている。しかし、間接被曝の健康影響については、線量の推定が困難であることや、おそらく健康影響を及ぼすような線量ではないであろうという理由から詳しく検討されてこなかった。しかし、近年になり原爆投下当日（8月6日）の入市被爆者における白血病が8日以降に入市した人に比べ有意に高かったという報告や、直接被曝による死亡危険度が西方向に有意に高くなっているという間接被曝による健康影響を示唆する報告がされている。

【目的】 本研究の目的は、入市被爆者のコホートデータを用いて、間接被曝による固形がんの死亡危険度を評価することにある。また、間接被曝の主な曝露源と考えられる放射化された核種の探索を行うことにある。

【資料および解析方法】 対象は、広島大学原爆被爆者データベースに登録され入市被爆者と認定され、1970/1/1で生存し2010/12/31まで追跡された男性27,062名、女性20,082名である。固形がんによる死亡をイベント、転出、他の病気での死亡あるいは観察終了時点での生存を打ち切り症例とし、発がんの多段階仮説に基づいた数理モデルを用いて生存解析を行った。被曝時年齢を a 、暦年を y 、被爆線量を D とすると、到達

年齢が t での固形がんによる死亡危険度 h は、 $h(t|y, D) = c \left(1 + \frac{\beta D}{t}\right)^{k-1} e^{g(t, y)}$ と表さ

れる。ここで β は被爆線量効果を表す係数であり、 $ce^{g(t, y)}$ は線量0の場合の広島原爆被

爆者のベースラインハザードを表している． $g(t, y)$ は 1970 年から 2010 年までの日本全国の固形がんによる 5 歳階級別死亡危険度の対数値を，到達年齢 t ，暦年 y ，到達年齢と暦年の交互作用 $t \times y$ を用いて 5 次の多項式で表した関数である．曝露を受けた人の非曝露の人に対する相対危険度 rr は以下の式で表すことができる

$$rr(t | D, y_1, y_0) = \left(1 + \frac{\beta D}{t}\right)^{k-1} e^{g(t, y_1) - g(t, y_0)}.$$
 ここで y_1, y_0 はそれぞれ曝露を受けた人および非曝露の人の暦年を示す． $\beta_1 = (k-1)\beta / 75$ とすると，相対危険度の対数値は近似的に $\log rr(t | D, y_1, y_0) \cong \beta_1 D \frac{75}{t} + g(t, y_1) - g(t, y_0)$ と表すことができる．入市者の

場合，被爆線量の同定が困難であることから，被爆線量として入市日を放射線量の代替指標とし，入市日と死亡相対危険度の関係について，被爆線量の被爆時年齢依存性を考慮に入れた数理モデルを構築し，生存解析を行った．また，被爆線量を放射化された核種の半減期と曝露時間を用いて表すことにより，曝露源となり得る放射化された核種の探索を行った．

[結果] 8 月 6 日入市者は 8 月 9 日以降の入市者に比べ固形がんによる死亡相対危険度が有意に高かった．7 日，8 日入市者については死亡危険度の有意な上昇は認められなかった．また，8 月 6 日入市者については，中年の人がその他の年代の人に比べ死亡危険度が高かった．46 歳で 8 月 6 日に入市した男性が 75 歳になった時の死亡相対危険度は対照群と比較して 22%，49 歳で入市した女性は 69% 高かった．また主に 8 月 6 日の入市者の死亡危険度の上昇に寄与した主な曝露源として，半減期が 2.58 時間の ^{56}Mn が考えられた．

[考察] 8 月 6 日入市者の死亡危険度の被爆時年齢依存性について得られた結果に対して以下のようなことが考えられる．原爆炸裂当日，多くの人々が広島市周辺地域から家族や友人の消息を求めて入市したと思われる．特に中年の男女は爆心地域付近まで立ち入り，長時間いろいろな作業に携わった可能性が高く，放射化した物質や放射性降下物を含んだ塵等を多く吸い込んだ可能性が考えられる．間接被爆による健康影響を示唆する本研究の結果は，2011 年 3 月の福島第一原発の事故により被曝した福島県民や原発事故に携わった人々の今後の健康状態を長期間にわたり注意深く観察していく必要があることを示している．

科学研究費・基盤研究(A)「非対称・非線形統計理論と経済・生体科学への応用」研究代表者：谷口正信（早稲田大学基幹理工学部）によるシンポジウム「バイオ統計学の調整と貢献」

シドニー大学ジョージ国際保健研究所における大規模臨床試験

有馬 久富

滋賀医科大学アジア疫学研究センター

シドニー大学ジョージ国際保健研究所(The George Institute for Global Health, University of Sydney)

ジョージ国際保健研究所(The George Institute for Global Health)は、オーストラリア、イギリス、インドおよび中国の4か国において450名以上のスタッフを有するグローバルな研究所として、数多くの臨床試験や疫学研究を実施してきた。私は、2003年から2006年までVisiting Senior Research Fellow、2008年にはSenior Research Fellow、2009年から2011年までSenior Lecturer、2011年から2014年6月までAssociate Professorとしてシドニー大学ジョージ国際保健研究所に従事し、PROGRESS、ADVANCE、INTERACT2、ENCHANTEDなどの大規模臨床試験に関わってきた。今回は、ジョージ国際保健研究所の設立から現在までの経緯、実施されてきた大規模臨床試験の概要、大規模臨床試験における医療統計家の役割などについて報告する。

ジョージ国際保健研究所の前身である国際保健研究所(Institute for International Health)は、シドニー大学医学部のサポートのもと1999年に開設された。その後、イギリス、インドおよび中国にもオフィスが設立され、グローバルな研究所になるとともに、ジョージ国際保健研究所と改名された。現在では450名以上のスタッフを有し、世界50か国以上で数多くの臨床試験および疫学研究を実施している。”To improve the health of millions of people worldwide”が研究所のmissionである。

研究所設立当初の主要なプロジェクトは、PROGRESS試験およびBlood Pressure Lowering Trialists' Collaboration (BPLTTC)であった。PROGRESS試験は、脳血管障害二次予防における降圧療法の有用性を確立した多施設国際共同、無作為化二重盲検比較試験である(Lancet 2001;358:1033-1041)。BPLTTCは、1995年以降、降圧薬をもちいた大規模無作為化比較試験の結果を統合するメタ・アナリシスである(Lancet 2014;348:591-598など)。いずれのプロジェクトも国際保健機関(World Health Organization: WHO)および国際高血圧学会(International Society of Hypertension: ISH)の後援を得て実施されていた。私が企画からかかわった臨床試験としては、脳出血急性期における積極的降圧療法の有用性を検討したINTERACT1およびINTERACT2試験がある。INTERACT試験

では、頭部 CT で確定診断にいたった発症 6 時間以内の脳出血患者のうち収縮期血圧 150–220mmHg の対象者を、積極的降圧療法群(降圧目標：収縮期血圧 140mmHg 未満)あるいは AHA/ASA 脳出血治療ガイドラインに基づく通常治療群(180mmHg 未満)に無作為に割り付け、90 日間追跡した。404 名を対象とした INTERACT1 試験では、積極的降圧療法が有害事象を増やすことなく有効性の代替エンドポイントである血腫増大を予防することを明らかにした(Lancet Neurol 2008;7:391–399)。2839 名を対象とした INTERACT2 試験では、積極的降圧療法が死亡や有害事象を増加させずに身体機能を改善することを報告した(N Engl J Med 2013; 2013;368:2355–2365)。

ジョージ国際保健研究所には、academic な研究を行いながら臨床研究の統計学的側面を補助する医療統計学者および解析業務を行う統計家が在籍していた。INTERACT2 試験には、1 名の医療統計学者と 2 名の統計家に従事していただいた。医療統計学者および統計家とは、企画の段階から密接な連携をとり、プロトコルの作成をサポートしていただいた。統計家 1 名のみが非盲検とされ、中間解析結果を安全評価委員会に報告していただいた。非盲検とされた統計家および安全評価委員会を除くすべての研究者・医療統計学者・統計家が割付群別の結果を見ることができなくするよう、研究所の標準業務手順書に定められていた。試験が終了する 1 年ほど前から、盲検下にある研究者および統計学者が中心となって統計解析計画書が作成され、Int J Stroke 誌に公開された。非盲検とされた統計家は、計画書の通りに統計解析を行った。試験が終了し、データに変更を加えることができなくなった段階で、論文執筆委員に統計解析結果が開示され、論文執筆および学会発表の準備が行われた。学会発表の前日、試験に関わったすべての研究者に対する結果報告会が開催された。INTERACT2 試験の主解析の結果は、2013 年 5 月にロンドンで開催された European Stroke Conference の Late Breaking Clinical Trials セッションで発表され、その直後に論文が N Engl J Med 誌の website に公開された。

シドニー大学ジョージ国際保健研究所に在籍中、いくつかの大規模臨床試験に関わることができた貴重な経験を、今後の日本での研究に生かしたい。

処置前後に観察された連続 Bounded Outcome Scores に対する 2 変量反応モデル

大分大学医学部 数学・統計学 大山 哲司

Bounded outcome score (BOS) とは有限区間に制限されたスコアである．例えば visual analog scale (VAS) は被験者の精神状態を測定する際によく用いられるが，これは 100mm の水平な線分からなり，両端を測定項目の最も極端な状態としたものである．被験者は自分の状態に対応する箇所に主観的に印をつけ，左端からの距離が測定値として用いられる．このスコアは开区間 $(0, 100)$ では連続であるが，両端の 0 や 100 という値もとることが可能な閉区間のスコアである．

このように区間が制限されたデータの分布は J 字型や U 字型などの非正規性を示したり，分散不均一性を示したりすることがある．解析方法としてはノンパラメトリックな方法の適用が考えられるが，分布を仮定できる場合は検出力が低下してしまう．変数変換する方法もあるが，非正規性と分散不均一性の問題を同時に解決できるとは限らず，解決できたとしてもそれが複雑な変換であれば，元のスケールでの結果の解釈が困難になってしまうと考えられる．Kieschnick and McCullough [1] は，开区間 $(0, 1)$ で観察されるデータに対して適用可能ないくつかの回帰モデルで実際のデータを解析し，比較を行っている．その結果から，彼らはベータ分布を仮定した回帰モデルを推奨している．それとは別に，Ferrari and Cribari-Neto [2]，Smithson and Verkuilen [3] もまたベータ分布を仮定したベータ回帰モデルを提案している．

処置前後の 2 時点で連続な BOS が観察された場合，前後差をとってから t 検定により群間比較を行ったり，共分散分析で解析されたりすることが多い．しかしこれらは正規性に基づく手法であるため，連続 BOS の処置前後変化を調べる際には適切でない．変数変換した上で解析することもあるが，1 時点の場合と同様に非正規性と分散不均一性の問題を同時に解決できるとは限らない．Cepeda-Cuervo et al. [4] は 2 変量ベータ分布を仮定した 2 変量ベータ回帰モデルを提案しているが，相関が比較的弱い範囲に制限されており，この範囲を超える相関を有するデータに対しては適用できない．

そこで本発表では，連続な BOS が処置前後の 2 時点で観察された場合の 2 変量モデルを提案した．まず， $U_j (j = 1, \dots, J)$ は測定 j における処置への反応強度を表す変数， W は反応をコントロールする被験者効果を表す変数であると仮定する．変数 W は，連続 BOS の観察において反応強度が同程度であっても控えめな反応を示す被験者もいれば，その逆の被験者もいるであろうといった，個人差を表す変数である．さらに，区間 $(0, 1)$ における連続な反応 Y_j は

$$Y_j = \frac{U_j}{U_j + W} \quad (1)$$

により観察されると仮定する．これにより，反応強度 U_j が大きいほどスコア Y_j も大きくなり，被験者効果 W が大きいほどスコア Y_j は小さめに抑えられることになる．次に， $U_1, \dots, U_J, W$ に対して以下のようにガンマ分布を仮定する．

$$U_j \sim \text{GAM}(a_j, b) \ (j = 1, \dots, J), \quad W \sim \text{GAM}(c, b) \quad (2)$$

ここで $a_1, \dots, a_J, b, c > 0$ である．このとき， $Y_1, \dots, Y_J$ の同時分布は Olkin and Liu [5] により与えられた多変量ベータ分布となる．特に $J = 2$ の場合，つまり 2 変量ベータ分布の確率密度関数は， Y_j の周辺平均

を μ_j として

$$f(y_1, y_2; \mu_1, \mu_2, c) = \frac{\Gamma\left(\left(\frac{\mu_1}{1-\mu_1} + \frac{\mu_2}{1-\mu_2} + 1\right)c\right)}{\Gamma\left(\frac{\mu_1}{1-\mu_1}c\right)\Gamma\left(\frac{\mu_2}{1-\mu_2}c\right)\Gamma(c)} y_1^{\frac{\mu_1}{1-\mu_1}c-1} y_2^{\frac{\mu_2}{1-\mu_2}c-1} \times (1-y_1)^{\frac{c}{1-\mu_2}-1} (1-y_2)^{\frac{c}{1-\mu_1}-1} (1-y_1 y_2)^{-\left(\frac{\mu_1}{1-\mu_1} + \frac{\mu_2}{1-\mu_2} + 1\right)c}. \quad (3)$$

と表すことができ、これが提案する 2 変量モデルのランダム成分である。被験者 $i (= 1, \dots, n)$ に対し処置前後の連続 BOS を (Y_{i1}, Y_{i2}) として、これがパラメータ $(\mu_{i1}, \mu_{i2}, c_i)$ の 2 変量ベータ分布 (3) に従うと仮定する。 μ_{ij} に関連する説明変数ベクトルを $\mathbf{u}_{ij}$, c_i に関連する説明変数ベクトルを $\mathbf{z}_i$ として、

$$g(\mu_{ij}) = \boldsymbol{\alpha}'_j \mathbf{u}_{ij} \quad (j = 1, 2), \quad h(c_i) = \boldsymbol{\gamma}' \mathbf{z}_i \quad (4)$$

というモデルを仮定する。ここで g, h は単調で微分可能なリンク関数である。適用例では μ_{ij} に対してロジットリンク, c_i に対して対数リンクを考える。さらに処置前後でのスコアの変化と、変化に影響する要因を調べるためにオッズ比型の尺度 $\psi_i = \frac{\mu_{i2}}{1-\mu_{i2}} / \frac{\mu_{i1}}{1-\mu_{i1}}$ を定義する。 $\psi = 1$ は処置前後でスコアの平均の変化がないことを意味し, $\psi > 1$ は処置後にスコアの平均が増加すること, 逆に $\psi < 1$ は処置後にスコアの平均が減少することを意味する。このとき ψ に対するモデルは

$$\log \psi_i = \text{logit}(\mu_{i2}) - \text{logit}(\mu_{i1}) = \boldsymbol{\beta}' \mathbf{x}_i \quad (5)$$

と表すことができる。ここで $\mathbf{x}_i$ はスコアの変化に関連する被験者 i の説明変数ベクトル, $\boldsymbol{\beta}$ は対応する回帰係数ベクトルである。未知パラメータの $(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\gamma})$, あるいは $(\boldsymbol{\alpha}_1, \boldsymbol{\beta}, \boldsymbol{\gamma})$ は最尤法により推定することができる。

上記の 2 変量モデルの適用例として, VAS を用いたオセルタミビル (商品名: タミフル) の臨床研究データ [6] を取り上げた。

参考文献

- [1] Kieschnick, R., McCullough, BD. (2003). Regression analysis of variates observed on $(0, 1)$: percentages, proportions and fractions. *Statistical Modeling* **3**, 193-213.
- [2] Ferrari, S.L.P., Cribari-Neto, F. (2004). Beta regression for modeling rates and proportions. *Journal of Applied Statistics*, **31**, 799-815.
- [3] Smithson, M., Verkuilen, J. (2006). A Better Lemon Squeezer? Maximum-Likelihood Regression With Beta-Distributed Dependent Variables. *Psychological Methods*, **11**, 54-71.
- [4] Cepeda-Cuervo, E., Achcar, J.A., Lopera, L.G. (2014). Bivariate beta regression models: joint modeling of the mean, dispersion and association parameters. *Journal of Applied Statistics*, **41**, 677-687.
- [5] Olkin, I., Liu, R. (2003). A bivariate beta distribution. *Statistics and Probability Letters*, **62**, 407-412.
- [6] Suzaki, Y., Uemura, N., Takada, M., Ohshima, T., Itohda, A., Morimoto, T., Imai, H., Hamasaki, H., Inano, A., Hosokawa, M., Tateishi, M., Ohashi, K. (2013). The effect of *carboxylesterase 1* (*CES1*) polymorphisms on the pharmacokinetics of oseltamivir in humans. *European Journal of Clinical Pharmacology*, **69**, 21-30.

がん第Ⅰ相試験における2剤併用療法の用量探索法：最近の展開

平川晃弘

名古屋大学医学部附属病院 先端医療・臨床研究支援センター
生物統計学・バイオインフォマティクス分野
hirakawa@med.nagoya-u.ac.jp

本稿では、2剤併用療法のための用量探索法に焦点を当て、近年提案された3つの方法を取り上げ、その方法論上の特徴を整理した。また、シミュレーション実験によりその性能を評価した。3方法は全てモデルに基づく用量探索法であるが、そのデザイン上の特徴はそれぞれが非常に固有的在であると言える。YY法は、各薬剤の用量ごとに毒性確率の当て推量を指定すればよいから、各薬剤の単剤試験での毒性発現確率を直接活用できる。他方で、WCO法は全用量組合せの当て推量が必要であり、単剤試験を含め既存の知見から予想する必要がある。それゆえ、当て推量が真値と大きく乖離する可能性も高く、この場合には性能が著しく低下する可能性がある。HHM法は当て推量が不要であり、事前に2剤併用に関する情報が全くない場合に有用かもしれない。また、YY法はモデルパラメータに対する事前分布の指定が必要であり、無情報事前分布を仮定したとしても、その感度分析が試験実施前に必要であると考えられる。コホートサイズと用量移動制限については、現時点でそのコンセンサスがないため、著者によって異なっている。ただし、2剤同時増量や用量スキップについては、安全性の観点から許容されない場合もあるため、実施する試験に適したコホートサイズと用量移動制限を用意すればよいと考える。

次に、シミュレーション結果に基づいて、3つの方法の性能を評価する。HHM法とWCO法のMTDC選択割合は他の2方法よりも高く、その有用性が示唆されたと言える。YY法についてはMTDC選択割合が極端に低くなる場合があることに留意が必要である。過毒性用量選択割合については、HHM法が最も低く、これに伴い毒性発現割合も最も低かった。本稿の目的は4方法の性能比較ではないので、これ以上の検討は実施しないが、性能比較の観点から言えば、少なくともコホートサイズ、用量移動制限を統一した下で、その性能を比較する必要があるだろう。また、性能評価の主たる指標として、MTDC選択割合ではなく、目標毒性確率の周辺用量（例えば、毒性確率20~40%）の累積選択割合を採用する場合もある。2剤併用療法の性能を1つの指標で評価することは困難であり、多角的な評価が必要である。ここで示したのは代表的と考えられる12シナリオに限定したシミュレーション結果であり、他のシナリオも含めて、各方法の相対関係を結論付けることには限界があることに留意されたい。抗がん剤の用量探索法については、抗がん剤開発の趨勢に伴い発展を遂げてきたと言える。単一の細胞障害性薬剤による治療が主流だった時代にCRMが提案され、多剤併用療法による治療が増加してきた現在では、これらに適した新しい用量探索デザインが開発されてきた。他方で、分子標的薬の台頭により、有効性と毒性を同時評

価する用量探索法が必要になり、こちらについても新しい用量探索デザインが数多く開発されている(大門, 2012). 例えば, 開発の早期において何らかのバイオマーカーを有効性の指標として利用可能な場合は, バイオマーカーと毒性を同時評価する用量探索法を利用できる. バイオマーカーを 2 値化 (例えば, 反応あり又はなし) できる場合は, Thall et al. (2004), Braun (2002), Zhang et al. (2006)などが利用可能である. バイオマーカーを連続量のまま利用する場合は, Bekele & Shen (2005)や Hirakawa (2012)などが利用可能である. ただし, 開発の早期において利用可能なバイオマーカーが同定されていることは稀であり, 毒性のみを指標とした用量探索が実施されているが現状である (Daimon et al., 2014)

昨今の個別化医療の加速に伴い, 早期臨床試験のあり方も再検討されている. 例えば, 開発早期にバイオマーカー等の有効性に関する指標を同定するために, 非臨床試験と同時並行で Phase 0 試験が実施されることがある. Phase 0 試験では, 対象疾患患者 2~3 例に対してごく低用量を短期間投与し, 薬物動態を含めた有効性の予備的評価やバイオマーカー探索が実施される (Rubinstein, 2012). 他にも, 通常の第 I 相試験成績に基づき, 有効性が期待できるサブ集団を特定し, それらの症例を同一の第 I 相試験に追加登録するデザイン (拡大コホートを用いた第 I 相試験) も注目されている (Manji et al., 2013). 今後はバイオマーカー陽性例と陰性例ごとに MTD を探索するデザインも必要になると考えられる.

参考文献

- Bekele, B. N. and Shen, Y. (2005). A Bayesian approach to jointly modeling toxicity and biomarker expression in a phase I/II dose-finding trial. *Biometrics* **61**, 344-354.
- Braun, T.M. (2002). The bivariate continual reassessment method: extending the CRM to phase I trials of two competing outcomes. *Controlled Clinical Trials* **23**, 240-256.
- 大門貴志 (2012). 連続再評価法とその周辺(I) -連続再評価法とは-. 計量生物学 **33**, 1-29.
- Hirakawa, A. (2012). An adaptive dose-finding approach for correlated bivariate binary and continuous outcomes in phase I oncology trials. *Statistics in Medicine* **31**, 516-532.
- Manji, A., Brana, I., Amir, E., Tomlinson, G., Tannock I. F., Bedard P. L., Oza, A., Siu, L. L., Razak A. R. (2013). Evolution of clinical trial design in early drug development: systematic review of expansion cohort use in single-agent phase I cancer trials. *Journal of Clinical Oncology*: doi :10.1200/JCO.2012.47.4957.
- Rubinstein, L. Statistics of Phase 0 Trials. In Crowley, J. and Hoering, A. (ed). (2012). *Handbook of Statistics in Clinical Oncology, 3rd edition*. New York: Chapman and Hall/CRC Press.
- Thall, P. F. and Cook, J. D. (2004). Dose-finding based on efficacy-toxicity trade-offs. *Biometrics* **60**, 684-693.
- Zhang, W., Sargent, D.J., and Mandrekar, S. (2006). An adaptive dose-finding design incorporating both toxicity and efficacy. *Statistics in Medicine* **25**, 2365-2383.

Multiple Imputation 法のモデル選択規準

野間久史（統計数理研究所），五所正彦（筑波大学医学医療系）

Multiple Imputation 法は，臨床研究・疫学研究における不完全データの統計解析において最も広く利用されている方法論であり，昨今の臨床試験における欠測データの取り扱いについての議論を受けて（例えば，Little *et al.*, 2012），実践においてもますます重要な方法論となっている．Multiple Imputation 法における実践上の最も悩ましい問題のひとつとして，補完値の妥当性（換言すれば，補完値の生成モデルの妥当性）の評価がある．Rubin (1987) による正則（proper）な補完方法のもとでは，単純には，ベイズ流のモデル評価方法を直接利用することができるが，一般的な統計ソフトウェアでは，マルコフ連鎖モンテカルロ法を利用した正則な補完方法は，限定的なモデルのもとでしか利用可能ではなく（例えば，SAS ver 9.4 では，proc mi が対応しているのは，多変量正規分布の構造が仮定できる場合のみである），実践的にはむしろ，回帰補完法や予測平均マッチングのような非正則（improper）な補完方法のほうが広く利用されている．また，近年の研究により，非正則な補完方法のもとの Multiple Imputation 法の頻度論的な評価は，Wang and Robins (1998), Robins and Wang (2000) などによって与えられている．

本研究では，これらの正則・非正則な補完方法を含めた，一般的な枠組みのもとで，Multiple Imputation 法のモデル選択規準を開発する．開発された規準は，一般的な MAR (Missing At Random) の仮定のもとのモデル評価は当然のこと，NMAR (Not Missing At Random) のもとでのパターン混合モデルの解析などにも利用することができ，パラメータの識別可能性を担保するための種々の付加的なモデルの仮定を評価することも可能であり，さらに MAR のもとでの解析モデルとの比較を行うことも可能であることを示す．

参考文献

- Little, R. J., D'Agostino, R., Cohen, M. L., *et al.* (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine* **367**, 1355-1360.
- Robins, J. M., and Wang, N. (2000). Inference for imputation estimators. *Biometrika* **87**, 113-124.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.
- Wang, N., and Robins, J. M. (1998). Large sample theory for parametric multiple imputation procedures. *Biometrika* **85**, 935-948.

2 群判別のための多重マーカーの一般化 t-統計量について： セミパラメトリックスとラッソ

江口 真透・小森 理 (統計数理研究所)

判別解析において最も知られている方法の一つは Fisher (1936) まで遡る。最近、機械学習の文脈でサポートベクターマシン、ランダムフォレストなど斬新な方法が提案されているがフィッシャー線形判別はしばしば、これらの方法と同程度の性能が報告されている。これはフィッシャー線形判別のアルゴリズム的な簡明さを考えると特筆すべきことだと言える。 p 次元の変数ベクトル $\mathbf{X}$ が 2 群の母集団 $Y = 0, 1$ から観測され、正規性と等分散性の仮定：

$$\mathbf{X}|(Y = y) \sim N_p(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_0) \quad \text{for } y = 0, 1 \quad (1)$$

をすると対数尤度比は定数を除いて線形判別関数 $\boldsymbol{\beta}_0 \mathbf{x}$ となる。ここで

$$\boldsymbol{\beta}_0 = \frac{\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}{\{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\}^{\frac{1}{2}}}. \quad (2)$$

この $\boldsymbol{\beta}_0$ に対するもう一つの導出は標準化された 2 群の平均差の最大化によって得られる：

$$\boldsymbol{\beta}_0 = \operatorname{argmax}_{\boldsymbol{\beta} \in \mathbb{R}^p} \mathbb{L}_I(\boldsymbol{\beta}), \quad (3)$$

ここで

$$\mathbb{L}_I(\boldsymbol{\beta}) = \frac{\boldsymbol{\beta}^\top (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}{(\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_0 \boldsymbol{\beta})^{\frac{1}{2}}}. \quad (4)$$

この $\mathbb{L}_I(\boldsymbol{\beta})$ を単変量 t-統計量の類推から、単に t-統計量と呼ぶ。

本論では $Y = 0$ 群の $\mathbf{X}$ の分布は正規性を仮定するが $Y = 1$ 群の $\mathbf{X}$ の分布は特定の分布を仮定しないセミパラメトリックな設定を考える。概要で述べられたようにケース・コントロール研究などを念頭に置いて、2 群の母集団モデルに対して、このような非対称な仮定のもとで次のような (4) の拡張を考える：

$$\mathbb{L}_U(\boldsymbol{\beta}) = \mathbb{E}_1 \left\{ U \left(\frac{\boldsymbol{\beta}^\top (\mathbf{X} - \boldsymbol{\mu}_0)}{(\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_0 \boldsymbol{\beta})^{\frac{1}{2}}} \right) \right\}, \quad (5)$$

ここで U は (フィルターの役目をする) 生成関数で、 $\mathbb{E}_1$ は $Y = 1$ を与えたときの条件付き期待値を表す。Cf. Komori et al. (2015). このように U の選択によって様々な効用関数 $\mathbb{L}_U$ が考えられる。 U が恒等関数ならば $\mathbb{L}_U$ は (4) の $\mathbb{L}_I$ に帰着される。 U が標準正規分布関数 Φ ならば、正規性の仮定から

$$\Phi \left(\frac{\boldsymbol{\beta}^\top (\mathbf{x} - \boldsymbol{\mu}_0)}{(\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_0 \boldsymbol{\beta})^{\frac{1}{2}}} \right) = \mathbb{P}(\boldsymbol{\beta}^\top \mathbf{X}_0 < \boldsymbol{\beta}^\top \mathbf{x} \mid \mathbf{x}), \quad (6)$$

なので

$$\mathbb{L}_\Phi(\boldsymbol{\beta}) = \mathbb{E}_1\{\mathbb{P}(\boldsymbol{\beta}^\top \mathbf{X}_0 < \boldsymbol{\beta}^\top \mathbf{X}_1 \mid \mathbf{X}_1)\} = \mathbb{P}(\boldsymbol{\beta}^\top \mathbf{X}_0 < \boldsymbol{\beta}^\top \mathbf{X}_1) , \quad (7)$$

となる．ここで $\mathbf{X}_y$ は $Y = y$ を与えたときの $\mathbf{X}$ の条件付き確率ベクトルとする．従って $\mathbb{L}_\Phi(\boldsymbol{\beta})$ は線形判別関数 $\boldsymbol{\beta}^\top \mathbf{x}$ の受信者動作特性カーブの下側面積 (AUC) となる．AUC は判別性能を測る標準的な尺度である (Pepe, 2003)．

2 群のサンプルサイズを n_0 と n_1 とし、 $Y = 0$ における正規性の仮定から標本平均 $\bar{x}_0$ と標本分散行列 $\mathbf{S}_0$ を採る．提案される一般化 t-統計量による推定量は

$$\hat{\boldsymbol{\beta}}_U = \operatorname{argmax}_{\boldsymbol{\beta} \in \mathbb{R}^p} L_U(\boldsymbol{\beta}) , \quad (8)$$

と与えられる．ここで $L_U(\boldsymbol{\beta})$ は $\mathbb{L}_U(\boldsymbol{\beta})$ の自然なサンプル版で

$$L_U(\boldsymbol{\beta}) = \frac{1}{n_1} \sum_{j=1}^{n_1} U\left(\frac{\boldsymbol{\beta}^\top (\mathbf{x}_{1j} - \bar{\mathbf{x}}_0)}{(\boldsymbol{\beta}^\top \mathbf{S}_0 \boldsymbol{\beta})^{1/2}}\right) . \quad (9)$$

とする．以下の節で推定量 $\hat{\boldsymbol{\beta}}_U$ のクラスの中で漸近一致性と漸近正規性を調べて、セミパラメトリックモデルにおいて最適な生成関数 U を求めた．さらに、推定量 $\hat{\boldsymbol{\beta}}_U$ を求めるアルゴリズムを与えて、説明変数 $\mathbf{X}$ が高次元の場合においてラッソ・アルゴリズムも検討した．

References

- [1] Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association* **70**, 892–898.
- [2] Eguchi, S. and Copas, J. (2002). A class of logistic-type discriminant functions. *Biometrika* **89**, 1–22.
- [3] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179–188.
- [4] Goeman, J. J. (2010). L_1 penalized estimation in the cox proportional hazards model. *Biometrical Journal* **52**, 70–84.
- [5] Komori, O., Eguchi, S. and Copas, J. (2015). Generalized t-statistics for two-group classification. To appear in *Biometrics*.
- [6] Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, New York.
- [7] Su, J. Q. and Liu, J. S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association* **88**, 1350–1355.

脳画像データ解析のための RBF-sPLS の適用

吉田寿子^{a,d}, 川口淳^b, 山下典生^c, 鶴屋和彦^d

はじめに

アルツハイマー病 (AD: Alzheimer's disease) は認知症の病型のひとつであり, 発症の原因としては, 遺伝的要因のみでなく, 生活習慣などの環境的誘因が複雑に関与していると考えられている. 我々はこれまでに, 核磁気共鳴画像法 (magnetic resonance imaging, MRI) データを用いて, 脳萎縮と関連する臨床症状を明らかにするための radial basis function-sparse partial least squares (RBF-sPLS) という手法を提案した^{1,2)}. 提案した RBF-sPLS は, PLS 回帰の前処理として, 脳画像の次元縮小に動径基底関数 (RBF) を用いているが, 本研究ではさらに, ネットワークに基づくクラスタリングを非画像バイオマーカーの前処理として取り入れ, 説明変数の次元縮小を行いつつ, クラスターの情報を結果の解釈に活かすという network based RBF-sPLS というモデルへ発展させる. このことによってアルツハイマー病に関連するバイオマーカーと脳萎縮領域を同時に明らかにすることを目指す.

2. 対象とデータ

AD に関する大規模データベースである the Alzheimer's Disease Neuroimaging Initiative (ADNI) の登録データを使用する. 具体的には, 脳 MRI のほか, 非画像バイオマーカーとして, apoE ε4 アリル, アミロイド β やタウ,

アポリポプロテインなどの脳脊髄液マーカーのほか, 末梢静脈血より採取された生化学・全血算データ, 身長・体重, および血圧などのバイタルサインデータを使用した.

3. 方法

まず, 血液や CSF バイオマーカーを含む非画像データに対し, ネットワークに基づいて重複を許すクラスタリングを行なった. ここで作成された p 個の各クラスターから代表変数を抽出し, sPLS の $n \times p$ の説明変数行列 $\mathbf{X}$ とした.

次に, sPLS に用いる N 次元の脳画像データ ($\mathbf{Y}_0$) の前処理として, RBF による次元縮小を行った. ここでは, 下記のように定義される³⁾動径 B-スプライン基底関数 $\phi(\cdot)$ を考える.

$$\phi(d) = \frac{1}{4g^2} \begin{cases} g^3 + 3g^2(g-d) \\ + 3g(g-d)^2 - 3(g-d)^3, & (d \leq g), \\ (2g-d)^3, & (g < d \leq 2g), \\ 0, & (d > 2g), \end{cases}$$

d は画像上の座標に等間隔に配置されたノットとボクセルとの距離, g は対角に隣接するノット間の距離であり, ボクセルがノットから離れるほど, 動径的に $\phi(d)$ は小さくなる. 対象の脳画像情報を $\mathbf{Y}_0 = (\mathbf{y}_{01}, \dots, \mathbf{y}_{0n})^T$, さらに, $b = \phi(d)$ を $N \times q$ 行列 $\mathbf{B}$ とおくと sPLS の従属変数 $\mathbf{Y}$ は次のように表記することができる.

$$\mathbf{Y} = \mathbf{Y}_0 \mathbf{B}$$

a. 久留米大学医学研究科 バイオ統計学, b. 京都大学大学院医学研究科 医学統計生物情報学,
c. 岩手医科大学医歯薬総合研究所 超高磁場 MRI 診断病態研究部門,
d. 九州大学大学院 包括的腎不全治療学

つまり、基底関数行列 $\mathbf{B}$ によって、MRI の従来の画素数 N は q 次元の行列 $\mathbf{Y}$ に縮小される。このようにして作成された $\mathbf{X}$ と $\mathbf{Y}$ に sPLS を適用し、診断名情報とは独立に 10 成分(脳萎縮領域とその領域に関連する非画像バイオマーカーのセット)を抽出した。その後、従属変数側(脳 MRI 側)のスコアを予測変数、登録時の診断名をアウトカムとするステップワイズ多項ロジスティック回帰モデルを用い、AD に関連するスコアのみを選択した。この手順によって選択されたスコアの属する成分が、AD に関連する脳萎縮領域と非画像バイオマーカーの組合せである。

4. 結果

本研究の解析には、2013 年 3 月時点で ADNI に登録されていた 819 例のうち、1.5T 脳 MRI 画像および非画像データのスクリーニングと欠損値補完を行い、534 症例(NC+MCI 430 例、AD 104 例)の登録時データを用いた。次元縮小によって得られた 7,176(次元)の脳画像データ $\mathbf{Y}$ とクラスタリングによって得られた 165 変数の非画像バイオマーカー行列 $\mathbf{X}$ に、sPLS を適用し、さらに成分を選択した結果、3 成分が AD に関連する脳領域とバイオマーカーの組み合わせとして選択された。最も AD との関連が強い成分は、萎縮領域として左右の海馬を含む脳領域と、apoE ε4 アリルを含む、リン酸化タウや総タウ、ヒト心臓由来脂肪酸結合蛋白などから構成されていた。これらは各クラスターの代表変数であり、この下位変数として少なくとも 1 つ以上のバイオマーカーが結合している。これらの下位変数を含めて、AD に関連する可能性のあるバイオマーカーとして解釈した。

5. 考察

本研究では、ADNI に登録された脳 MRI 画像と非画像バイオマーカーのデータを用いて、高次元データへの network based RBF-sPLS の適用について述べた。この手法では、多変量解析である sPLS の前処理として、画像データには RBF を、非画像データにはネットワーククラスタリングを用いてそれぞれ次元縮小を行なった。その結果、変数単位で結果の解釈を行なうのではなく、変数の集合を脳領域や、検査値が示す臨床的な特徴として解釈できるようになるというメリットが得られた。また、sPLS には直接利用されない変数も、結果の解釈には活かすことができる可能性が示唆された。欠損値の扱いや、各種パラメータの決定方法について考慮し、さらに高次元のデータへの適用事例⁴⁾でも高い精度を得られるよう検討を重ねたいと考える。

6. 文献

- 1) Yoshida H, Kawaguchi A, Tsuruya K., 2013, Radial basis function-sparse partial least squares for application to brain imaging data. *Comput Math Methods Med*, 2013:591032 (May)
- 2) 吉田寿子, 川口淳, 山下典生. 2013. 脳画像データとバイオマーカーとの関連解析 -sparse PLS の応用-日本統計学会誌 43(1), 85-97
- 3) Saranli A, Baykal B., 1998. Complexity reduction in radial basis function (RBF) networks by using radial B-spline functions. *Neurocomputing* 18(1-3), 183-194
- 4) 川口淳; 2015.構造的脳画像を目的変数とする方法. 2015 年度日本計量生物学会年会予稿集

次世代シーケンスデータによる体細胞変異・高多様性領域の解析

井元 清哉

東京大学医科学研究所ヒトゲノム解析センター

imoto@ims.u-tokyo.ac.jp

近年の次世代シーケンサーの急速な高性能化・低コスト化に伴い、大量のゲノム・トランスクリプトームシーケンスデータが計測され、その統計解析はがん研究をはじめとするさまざまな生命科学分野において重要な位置を占めている。本講演では、がん研究におけるゲノムシーケンスを用いた標準的な解析として体細胞変異の同定を取り上げる。がん細胞と同一患者の末梢血のシーケンスデータを比較することにより、原理的には後天的に獲得しがん細胞となった原因のゲノム変異（一塩基変異、挿入、欠損など）を同定することができる。一度に 100 塩基程度の長さでしか塩基配列を決定できないシーケンサー自体の原理的な限界やシーケンスエラーを統計処理するために全ゲノムシーケンスでは同一箇所を 30 から 40 回繰り返しシーケンスしたデータが得られる。全エクソームシーケンスでは 100 から 200 回程度の繰り返し標準的である。もし、シーケンスエラーがゲノム上どの箇所でも均一に起こり、がん細胞が均一（正常細胞の混入がなく、single clone のがん細胞）であれば体細胞変異の同定はそれほど困難な問題ではない。しかしながら、がんは均一ではなくシーケンスデータには正常細胞由来のものも混入する。従って、より正確に検出力の高い体細胞変異同定手法が望まれる。講演では、生殖細胞系に有するヘテロ接合性変異の情報を利用した体細胞変異同定のベイズ統計的方法論を紹介する。

参考文献

Usuyama, N., Shiraishi, Y., Sato, Y., Kume, H., Homma, Y., Ogawa, S., Miyano, S., Imoto, S. (2014) HapMuC: somatic mutation calling using heterozygous germline variants near candidate mutations, **Bioinformatics**, 30(23), 3302-3309.

わが国の生活習慣病とその合併症の現状と課題：久山町研究

九州大学大学院医学研究院附属総合コホートセンター

二宮利治

現在わが国は、国民の4人に1人が65歳以上という超高齢社会を迎えており、今後も早いスピードで高齢者人口が増加し、2050年には3人に1人が65歳以上になると推計されている。このような急速な人口の高齢化と生活習慣の欧米化に伴い、わが国では疾病構造の変化し、生活習慣病やその合併症などの慢性疾患対策が重要となった。なかでも、心血管病や認知症は生活の質、日常生活動作を阻害する最大の要因であり、その予防対策を講じることはわが国の医療行政における焦眉の課題である。そのためには、疫学研究によってわが国の生活習慣病とその合併症の実態を把握し、その危険因子を明らかにすることが不可欠であろう。本講演では、福岡県久山町で長年にわたり継続中の疫学調査（久山町研究）の成績を紹介し、わが国の地域住民における生活習慣病とその合併症の現状と今後の課題について検討する。

久山町は、福岡市の東に隣接する人口約8,000人の比較的小さな町である。この町の住民の年齢・職業構成は、調査開始時から現在に至るまで日本の平均レベルにあり、栄養摂取状況も国民健康・栄養調査の成績とよく一致している。つまり、久山町住民は日本人の標準的なサンプル集団と考えられる。この町では、1961年の研究開始時より、年代毎に40歳以上の地域住民を対象としてコホート集団を設定し、前向きに追跡調査を行っている。いずれの集団も当該年齢人口の約80%を調査対象とし、徹底した追跡調査システムより追跡脱落率が0.2%以下と極めて少ない。心血管病や死亡などのエンドポイントの発症は、病歴、理学的所見、臨床データ、画像所見を用いて診断を確定している。さらに、全町死亡例の約75%を剖検し、死因や臓器障害を病理学的に検証している。このような地域住民を対象に精度の高い調査を長期間継続した疫学研究は他に類

をみない。

久山町研究では、これまでに心血管病とその危険因子の時代的变化を明らかにした。1961年、1974年、1983年、1993年、2002年の集団をそれぞれ7年間追跡した成績を比較すると、脳卒中発症率は男女とも顕著に低下したが、男性では最近その低下が鈍化している。一方、虚血性心疾患発症率は、男女とも有意な変化を認めなかった。さらに、心血管病の危険因子の時代的变化を検討した。心血管病の強力な危険因子である高血圧の頻度は、1961年から2002年にかけて変化を認めなかったが、降圧薬服用者の頻度は有意に増加した。そのため、高血圧者における収縮期および拡張期血圧の平均値は、男女ともに有意に低下した。このような血圧管理の改善が、脳卒中発症率が時代とともに低下した大きな要因と考えられる。一方、肥満、高コレステロール血症、糖代謝異常といった代謝性疾患が、近年大幅に増加し、心血管病の重要な危険因子として台頭している。

また、久山町ではこれまで1985年、1992年、1998年、2005年、2012年に65歳以上の高齢住民を対象とした認知症の有病率調査を行った。各調査の受診率はそれぞれ90%以上といずれも高い。全ての調査で神経心理テストを用いてスクリーニング調査を行い、病歴聴取と神経・理学的所見、画像診断に基づき認知症の有無、病型を正確に診断している。これらの調査の結果、65歳以上の地域高齢者における認知症の有病率は、1985年から2012年にかけて6.7%から17.9%と急増した。さらに、認知症発症の危険因子や防御因子の探索を行ったところ、中年期高血圧、糖代謝異常（特に食後高血糖）、喫煙、遺伝的因子が認知症発症の有意な危険因子であり、豆腐、野菜、牛乳を中心とした食習慣や定期的な運動習慣が有意な防御因子であった。認知症を予防するためには、高血圧や糖尿病などの危険因子を管理するとともに、食生活や運動習慣を是正することが重要といえよう。

科学研究費・基盤研究(A)「非対称・非線形統計理論と経済・生体科学への応用」
研究代表者: 谷口正信(早稲田大学基幹理工学部)によるシンポジウム
「バイオ統計学の挑戦と貢献」

開催責任者: 米本孝二(久留米大学バイオ統計センター)

日時: 2015 年 2 月 2 日(月)～2 月 4 日(水)

場所: 九州大学医学部百年講堂中ホール3

タイトル: オーダーメイド医療研究の最前線

所属・氏名:

理化学研究所・統合生命医科学研究センター 副センター長 久保 充明

抄録:

2003 年のヒトゲノム解読完了を契機に、ハップマッププロジェクト、1000 人ゲノムプロジェクトなどヒトゲノムに関する国際的な研究基盤整備が進められるとともに、ヒトゲノムの解析技術が急速に進歩した結果、医学・医療に関するヒトゲノム情報の基礎的な知見が、近年、大量に蓄積されてきている。なかでも、多因子疾患に対するゲノムワイド関連解析の手法は、2002 年に理化学研究所から proof of concept が示され、2007 年以降、ゲノムワイド解析用 SNP アレイの商用化とともに世界中のゲノム医学研究の標準的解析手法となり、関節リウマチや炎症性腸疾患では 100 を超える疾患関連遺伝子が同定されるまでに至った。現在、世界中で、これまでに蓄積されたゲノム医学研究の知見や新たなゲノム解析技術を医療へ応用する試みが急速に進められており、特に欧米においては、SNP アレイやシーケンス技術の臨床応用を目指した研究が盛んに行われている。そもそも 1990 年に開始されたヒトゲノム計画自体が、ヒトゲノム情報の医療への応用を目的とした国際プロジェクトであり、これは自然な流れともいえる。我が国においても、平成 26 年 5 月に健康・医療戦略推進法および独立行政法人日本医療研究開発機構法が成立し、我が国を世界に先駆けて超高齢社会を迎える課題解決先進国として位置づけ、健康長寿社会の形成を目指した医療分野の研究開発に関する総合戦略が策定された。そこには再生医療と並んで「疾病克服に向けたゲノム医療実現化プロジェクト」が世界最先端の医療の実現に向けた取組として取り上げられており、個人のゲノム情報を用いた医療の実現に対する期待は大きい。

ヒトゲノム解読が完了した 2003 年に、文部科学省委託事業「オーダーメイド医療の実現プログラム(オーダーメイド医療実現化プロジェクト)」が開始された。本プログラムにおいては、第 1 期の 5 年間(平成 15～19 年度)に 12 医療機関の協力の下、20 万人、34 万症例、47 疾患の DNA・血清・臨床情報を有する世界最大規模の患者バンク「バイオバンク・ジャパン」を構築した。第 2 期の 5 年間(平成 20～24 年度)においては、「バイオバンク・ジャパン」を研究基盤とした大規模な疾患・薬剤ゲノム研究を実施し、病気のなりやすさや薬の効果・副作用に関係する遺伝子を多数発見し、これらの成果を Nature Genetics48 編を含む

268 編の論文として報告した(平成 26 年 10 月末時点)。これらの研究成果は、随時、プロジェクトのホームページで公開しているので、是非ご覧いただきたい。(オーダーメイド医療実現化プロジェクトホームページ:<http://biobankjp.org/>)

平成 25 年度より開始された第 3 期においては、研究基盤である「バイオバンク・ジャパン」を強化するため、がん・循環器疾患等を中心とした 38 疾患について新たに 10 万人を目標とした DNA・臨床情報の収集を開始するとともに、平成 26 年度には研究基盤の強化として血清・血漿バンクの増設および組織バンクの新設が決定した。これにより、バイオバンク・ジャパンは、DNA100 万本、血清・血漿 334 万本(液体窒素タンク 58 台)、凍結組織 48 万本(全自動型液体窒素タンク 12 台)を保管・管理可能な巨大なバイオバンクへと成長した。さらに、ヒトゲノム解析技術の進展に伴い、平成 25 年度より次世代型シーケンサーを導入し、バイオバンク・ジャパンに保管されている DNA サンプルを用いて全ゲノムシーケンス解析を開始した。また、平成 26 年 5 月には、第 1 期で収集された 20 万人について、ほぼ全例のゲノムワイド SNP 解析が終了し、20 万人、一人当たり 60~95 万 SNP のデータを有する日本人の疾患・薬剤関連遺伝子研究データベースを構築した。今後、全ゲノムシーケンスデータやゲノムワイド SNP データを種々の疾患・薬剤反応性情報と組み合わせて解析することにより、疾患易罹患性・薬剤反応性に関連する遺伝子の同定、薬剤開発や診断薬の標的分子の探索、遺伝・環境要因相互作用の探索を通じた重要な遺伝・医療情報の解明など、「オーダーメイド医療の実現」に向けたゲノム医学研究がさらに発展すると期待している。

本講演では、これまでのプロジェクトの歩みと本プロジェクトの最終目標である「我が国におけるオーダーメイド医療の実現」に向けた方向性について紹介する。最後に、本プロジェクトは多数の患者さん、ご家族、メディカル・コーディネーター、医療関係者、研究者など多くの関係者の多大なご協力により、支えられ、進められている活動であり、この場をお借りして、すべての関係者の方々に改めて感謝の意を表したい。

地方大学における臨床研究支援部門の取り組み ～統計家の視点から～

佐藤 俊太郎

長崎大学病院 臨床研究センター

shuntarosato@nagasaki-u.ac.jp

はじめに

長崎大学病院は、2000年に薬剤部内に治験支援センターを開設、2011年に「臨床研究センター」へ改組し、院内で実施される臨床研究の支援を開始した。2014年3月までは、臨床研究倫理委員会提出用の資料の作成支援（研究計画書、同意説明文書等）、UMIN登録、情報公開、必要に応じてモニタリングを実施してきた。2014年4月から生物統計家を配置し、統計学的事項への支援（研究計画書・論文作成時のアドバイス、統計学の勉強会）を始めた。

本稿では、2014年4月から集積を始めた臨床研究センター支援状況データベースの集計結果について報告し、勉強会の実施方法および地方大学の統計家が抱える問題を提起する。

臨床研究センター支援状況データベース報告

報告対象期間は、2014年4月1日～12月12日（175日；土日祝日除く）である。データベースは相談者の所属、相談（支援）内容等を集積している。集計のために相談内容は、1. 研究計画書作成、2. UMIN登録、3. 情報公開、4. モニタリング、5. 統計相談、6. その他の6項目に分類した。

期間中に226件の支援を実施し、統計相談は143件（63.3%）あった。統計相談は支援開始から順調に推移していた。特定の医局・部からの相談が多く（リピート率が高い）、断続的に宣伝を実施すると新規の医局・部・相談者からの相談が増加した。センターの支援内容は院内にまだ十分周知されておらず、認知されれば相談件数はさらに増加すると考えられる。

相談内容は、研究計画前段階で研究デザインや解析方法を相談するパターン、データを収集してから、もしくは解析を実施してから解析方法を相談するパターンの2つに大きく分けることができた。また相談内容の中には、統計学で全て解決できると誤解しているものも見受けられた。

統計勉強会の実施

毎月2回、統計初級者を対象にした勉強会を実施している。研究者からの要望もあり、途中から統計解析ソフトEZR（Rコマンドの改変）を用いた演習も実施している。勉強会の内容が少しずつ高度化するのに伴い、研究者の理解の仕方、勉強会に求める内容が二極化し、「初級者向け」ではなくなりつつある。これらの問題を受け、今後は初級者向けと

中級者向けの二つの勉強会を開催することを検討している。

地方大学の統計家が抱える問題

都市と地方の大学で研究者の数に違いはあると思うが、いずれにしても統計学的事項への需要・要求は多いと考えられる。特に地方大学ではその受け皿となる医学部・病院内所属の生物統計家の数が少ないのが現状である。長崎大学病院では諸問題についてディスカッションをする場が少なく、書籍、論文、Web、時にはシミュレーションで解決を図っている。また、医師主導治験や多施設共同前向き研究では、研究のシステム（統計解析も含む）に高い質が求められるが、ある程度の解析の質を担保できるダブルプログラミング、解析結果の確認を少ない人数では実施することができないといったシステム上の問題も生じる可能性がある。

結論

来年度の医学系研究の倫理指針の改訂に伴い、研究に求められるレベルが上がり、統計学的事項に対する需要・要求はさらに高まることが容易に予想できる。一方で前述の問題を抱える大学は多いと考える。統計家同士のネットワークをさらに強化し、質の高い人材の供給を増やすといった環境を改善する方法を考える必要がある。

IntCox は無増悪期間解析の Remedy になり得るか

東京慈恵会医科大学 西川正子
元久留米大学 森川敏彦

胃がんなどの臨床試験では、従来、主要評価項目は生存時間であったが、近年は、主要評価項目は生存時間ではなく無病生存時間(DFS, がんの再発又は死亡のいずれか早い方をイベントと定める)や無増悪生存時間(PFS, がん細胞の増悪又は死亡のいずれか早い方をイベントと定める)が多用されるようになった。患者の死亡前に再発や増悪が起こる場合が多く、がんが再発しているか否か、または患者のがん細胞が増悪しているか否かは検査して初めてわかる。したがって、がん細胞の増悪時間は増悪していない最後の検査時点から増悪が認められた最初の検査時点までの区間のある時点ということしか分からない。再発時間も同様である。このようにイベント発現までの時間が区間で与えられるようなデータを区間打ち切りデータと呼び、その区間を観察打ち切り区間と呼ぶ。一部の個体についてはイベント発現までの正確な時間が得られるが、そのほかの個体については区間打ち切りデータとして得られる場合は、特に「部分的区間打ち切りデータ」(partly interval-censored data, PICD)とよばれる。

PICD を用いて PFS を評価する場合、各被験者の増悪発現時間として区間打ち切りデータについては観察打ち切り区間の 1 点(例えば左端、中点、右端の値)を代入して既存の解析手法により治療効果を評価する簡便法と、観察打ち切り区間をそのまま用いて評価する方法(e.g. Turnbull (1976), Finkelstein (1986), Sun et al. (2005)). がある。一般には、1 点代入法が用いられ、中でも右端代入が慣例として用いられる場合が多い。右端代入法では、増悪時間の評価においては効果が過大評価になるが、右側中途打ち切りデータが存在する場合、1 点代入後に Kaplan-Meier 推定量を用いた集団としての無増悪生存率の推定では結果の挙動に不合理な点が見られること(e.g. Nishikawa and Tango (2003))、および Log-rank 検定では結果に不合理な挙動が見られること(Kawaguchi et al. (2012))などが判明している。また、1 点代入法では無増悪生存率の標準誤差が過小評価になり、観察打ち切り区間が広い場合や被験者によって異なる場合には推定値にバイアスが入ることが知られている。

生存時間解析において要因効果の大きさを定量的に検討する場合や、交絡因子についての調整を行って効果の大きさを定量的に検討する場合に通常最もよく用いられる回帰モデルは比例ハザードモデル(Cox の比例ハザードモデル)である。最近では 1 点代入をせずに観察打ち切り区間をそのまま用いて推定を行う方法も提案されている(e.g. Finkelstein (1986), Pan (1999)). Hu et al. (2012) はシミュレーションにより右端、および中点代入後に Cox の比例ハザードモデルをあてはめる方法(従来法)と Finkelstein 法(1986)をシミュレーションにより比較検討し、Finkelstein 法(1986)が

従来法よりも性能が良いことを示した。Pan (1999) の方法は Cox の比例ハザードモデルの区間打ち切りデータ対応版に相当し、R package の 'IntCox' として利用可能である。Chen et al. (2012) はシミュレーションにより区間打ち切りデータに対する 'IntCox' の性能を従来法と比較評価し、PFS の解析の Remedy として 'IntCox' の利用を推奨している。しかし、彼らのシミュレーションデータは区間打ち切りデータのみで、死亡のような正確なイベント発現時間は含まれていないし、シミュレーションの設定状況も通常の臨床試験ではあまり遭遇しない。そこで、本研究は PICD に対する 'IntCox' の性能を臨床試験でよく遭遇するような設定でシミュレーションにより比較評価した。また、PICD に対して 'IntCox' を用いる場合、必ずエラーや警告が返されるので、このような実用的な問題の対処法についても検討した。

1 点代入法についての理論的な検討として、生存分布の推定における打ち切りデータの影響をみるため、最も簡単な生存時間分布である指数分布を仮定し、打ち切りの影響を検討した。

シミュレーションは 2 つの治療法（抗癌剤など）又は予防法群の比較試験で、無増悪生存時間（PFS）を主要評価項目とした臨床試験を想定した。試験のフォローアップスケジュール、対照群の無増悪生存時間の真の分布およびデータの発生方法等の手順は、Nishikawa and Tango (2003) の臨床試験をもとにしたシミュレーションと同様な設定にした。真のハザード比（HR）は、1.0, 0.667, 0.5 を検討し、IntCox を用いる解析と、3 種類の一点代入（区間打ち切りデータに区間の左端、中点、右端をそれぞれ代入する方法）後に Cox の比例ハザードモデルをあてはめる方法とを比較評価した。シミュレーションの繰り返し数は 10000 回とした。

理論的検討の結論としては、考察の範囲内では右端代入は合理的根拠がなく、線形 1 次近似からは左端代入が、線形 2 次近似からは中点代入が推奨された。左端代入よりも中点代入が奨められる。シミュレーションでは、IntCox では 2 群の間に差がでにくい方向へのバイアスが一点を代入する従来法よりも大きかった（最も保守的）。非正常終了の場合では、IntCox のバイアスは従来法とほぼ同じであった。

本研究では、Chen et al. (2012) と異なるシミュレーション設定により 'IntCox' の性能を比較評価した。本研究では、PICD によりおもに右側打ち切りの影響を評価した。更なるバイオ統計学の挑戦が必要である。

この研究は日本学術振興会科学研究費助成事業（基盤研究（C））の援助を受けた。

ありふれた疾患の発症や血圧変動などの多くの複雑なヒト形質に、環境要因と共に一定の遺伝的要因が寄与することは間違いない。その遺伝的構成の理解は個別化医療・予防を実現する上で不可欠な知見である。しかし多くの遺伝的要因は未知な部分が多く、その構成の解明には、仮説に基づかない網羅的な探索が効果を発揮する。一塩基多型 (SNP) を用いたゲノムワイド関連研究 (GWAS) はゲノムワイドな探索を行う上で有用であり、多くの疾患の遺伝的要因が同定され、GWAS が用いられる分野はさらに拡大を続けている。一般的な GWAS では、各 SNP と形質との周辺的な関連性を調べる方法が標準的である。ところが同時に、それらが説明する形質の変動はあまり多くないこともわかってきた。SNP-GWAS データから推定された遺伝率 (例えば、Yang et al. (2010) の混合線形モデルによる推定) は、過去に双子研究のような系図研究から得られていた値よりも低く、失われた遺伝率の問題と呼ばれている (Manolio et al. 2009, Maher 2008)。血圧や2型糖尿病などの多因子性の遺伝的構成をもつと考えられている形質においては、標準的な SNP-GWAS を用いた形質値の予測はいまだ困難であり、実用に耐える水準には至っていない。

失われた遺伝率を読み解くために、様々な異なる接近法が模索されはじめている。代表的なものとしては、微小効果 SNPs、レアバリエント、遺伝子間相互作用、遺伝子 x 環境間相互作用、構造バリエントなどが挙げられる (Manolio et al. 2009)。これらの仮説を検証するための方法論は、まだ十分に整備されているとは言えず、今後のさらなる進展が待たれる。そのなかでも、遺伝子間相互作用、あるいは遺伝子 x 環境間相互作用は、既存の豊富な SNP-GWAS データ資源を直接利用することで直ちに探索可能となることがひとつの利点と言える。遺伝子間相互作用の存在は、過去の研究および生物学的な機序から多くの示唆がある反面 (Phillips 2008; Mackay 2014)、遺伝子 x 遺伝子間相互作用については、ゲノムワイドな探索からの有益な知見はまだ得られておらず、過小評価される傾向にあり、時にはそれが存在しないとまで評されることもある (Carlborg & Haley 2004; Crow 2010)。このような現状に対して、Moore は、「(遺伝子間相互作用の存在) が酵母で真実なら、なぜ、我々はヒトを酵母よりも単純と予想するのだろうか？」と警鐘をならしている (Eichler et al. 2010)。最近、系図ベースの狭義の遺伝率推定値が、ドミナンスや相互作用、共有環境によってインフレーションしていたと相次いで報告されている (Gaugler et al. 2014; Zaitlen et al. 2014)。それらはむしろ広義の遺伝率であるが、個人のリスク予測には、それらの非相加的成分や環境成分が重要になることを意味している。本研究では、SNP-GWAS データを用いてゲノムワイドに相互作用を探索する方法について検討する。SNP-GWAS は一般にサンプル数よりも候補要因数が遥かに大きくなる。例えば、30 万~100 万 SNPs に対して、サンプル数は数千程度が標準である。もし、加えて imputation が実施された場合、利用可能な SNP 数はさらに増加する。このような特性のデータを扱うための統計的枠組みは、近年になってようやく研究がはじまったものであり、数理統計学的にもまだ未熟である。例えば回帰モデルにおいては、 $p \gg n$ 問題と呼ばれ、近年活発に研究が進められている分野である。まだ多くの乗り越えるべき課題はあるが、Lasso などの高次元回帰法は人類遺伝学分野でもすでに導入されており、SNP-GWAS を用いた形質予測に関する応用も存在する (Wei et al. 2013, Abraham et al. 2013)。遺伝子間相互作用の探索において、2 次オーダーでの SNP 効果を調べる、所謂 SNP x SNP 間相互作用の解析は、最も容易な探索であろう。統計的交互作用項の検定は、相互作用の検出のためには統計的な意味では標準的である。SNP は 3 つのカテゴリをとるカテゴリ変数であり、その全 2 ペア (すなわち p 個の SNP に対して、 $p(p-1)/2$ 個のペア) について回帰モデルのもとで交互作用項の検定を行う。パラメータの導入方法には飽和モデルを含め複数の可能性がある。例えば、現行の SNP-GWAS で標準的な周辺回帰で用いられる各 SNP のアリル数に応じた加法モデルの自然な拡張から、加法モデルにおける交互作用項 (自由度 1) の検定がある (Cordell 2009)。ゲノムワイドな探索は SNP 数が大きいことで、計算負荷が増大し、そのため単純に全ペアについて統計量を計算するだけでもかなりの計算機リソースが要求される。例えば、ニュートン・ラフソン法のような繰り返し計算は計算負荷が大きく、ロジスティック回帰モデルの加法コーディングに基づく交互作用項の検定でもかなり高コストとなる。これらの計算を回避する方法がいくつか提案されている。例えば、PLINK に実装されている fast-epistasis 法は、アリル間の相互作用に基づく方法である。Wu et al. (2010) はハ

プロイド間の相互作用を、ロジスティック回帰モデルでの交互作用項の検定を用いて検出する方法を提案した。Ueki & Cordell (2012)は fast-epistasis 法や Wu らの方法などの SNP x SNP 間相互作用統計量の性質を調べ、それらの帰無分布が様々な相互作用以外の要因に影響を受けて偽陽性が生じることを見出し、それらの修正案を与えた。各遺伝子型の効果は常に加法的であるとは限らず、優性や劣性の遺伝様式がありうる。そのような形の相互作用が存在する場合、加法モデルは柔軟さに欠けるため、相互作用の構造を捉えきれず、検出力の低下を招くことが懸念される。その解決には、遺伝子型データに基づいた飽和モデルのもとで 4 つの交互作用項について検定（自由度 4）する方法が提案されている（Cordell 2009, Wan et al. 2010）。ただし自由度 4 の検定は、推定するパラメータ数が多いため、検出力が落ちるという点もある。また、観察される遺伝子型の分割表中に空セルが観察されるケースも生じうる。例えば、アリル頻度が低い場合や、連鎖不平衡などがその原因となる。Ueki (2014)は、観察された空セルを省いた適応的な自由度を提案した。

遺伝子間相互作用効果を検出する統計的手法はまだ確立されていない。例えば、仮説検定に基づく手法を用いたゲノムワイドな相互作用の探索は全 2 ペアの SNP の探索を行う場合、有意水準の設定は難しい問題である。もし、SNP-GWAS と同様に SNP 数でのボンフェローニ補正による多重検定補正を利用すれば、50 万 SNPs について、全 2 ペアの総数は約 10^{11} 個である。したがって、family-wise error rate を 5% とする場合、 $P < 5 \times 10^{-13}$ という非常に厳しいカットオフが、各 SNP x SNP 相互作用検定の仮説に適用されることとなる。これはあまりに厳しい要請であり、現状の GWAS でのサンプル数では十分な検出力を維持できない。Ueki & Tamiya (2012)は、Fan & Lv (2008)の sure independence screening (SIS)を用いた高次元変数選択法を応用した手法を提案した。SIS は、変数のランキングを行い、上位ランクの変数について変数選択法を適用する手法である。SNP x SNP 相互作用のランキングには、遺伝子間相互作用に由来から利用されていた MDR

(Multifactor Dimensionality Reduction, Ritchie et al. 2001) と類似の原理を用いている。そこで作成されたプログラム EPISIS は、GPGPU を利用した高速化により全 SNP ペアの網羅的計算を実現している。また、EPISIS は二値形質に対してのみ適用可能であったが、量的形質も扱える EPISIS の拡張版プログラム QEPISIS を開発した。

環境因子が測定されていれば、SNP-GWAS データから直ちに遺伝子 x 環境相互作用の探索が可能である。例えば、Hamza et al. (2011)が SNP の主効果と SNP x 環境効果の結合効果を調べる Kraft の結合効果検定 (Kraft et al. 2007) を用いることで、パーキンソン病におけるコーヒーと GRIN2A 遺伝子上の SNP の間での相互作用を報告している。ただし、統計的交互作用の検定を用いた場合の P 値はゲノムワイド有意水準を越えておらず、そのため統計的な意味で標準的な統計的交互作用の検定からは見いだせなかった効果である。Kraft の検定は、SNP の主効果と SNP x 環境相互作用を統計的に区別せず、まとめて検定を行うことで検出力を高めている。したがって、得られた結果の解釈、特に、真に相互作用効果であるかどうか、については生物学的な機能面からの注意深い検証が必須である。しかしながら、Hamza らの結果は、SNP-GWAS データから相互作用を捉えるためには、柔軟な統計手法を用いる戦略が利益をもたらす可能性があることを示唆している。Ueki & Tamiya (2012)の方法は、環境因子がカテゴリ変数であれば、遺伝子 x 環境相互作用について拡張できる。その手法はすでにプログラム EGSIS に実装済である。ところで、Hamza らの結果はケース・コントロール研究からのものであった。しかしながら、環境因子については、ケース・コントロール研究や横断研究などを用いた場合に、逆因果性が生じる危険性がある。したがって、データのサンプリングについては注意深いデザインが求められる。前向きコホートは、逆因果性をある程度除去することができるため有用であり、その上で環境因子を SNP-GWAS と組み合わせて利用すれば、結果の信頼度を高められるものと期待できる。このような研究デザインはゲノムコホートと総称され、英国での UK バイオバンクやアイスランドコホート、わが国では九州久山町コホートを嚆矢に、山形コホートなどが該当する。