特集 2　精神科領域の個別化医療の実現可能性とバイオバンキングの活用

2．精神科領域の個別化医療の実現可能性とバイオバンクの活用

高山　順[1,2]　田宮　元[2,3]

抄録：精神疾患をはじめとした複雑な疾患に対し，その遺伝要因はゲノム科学的手法により，環境要因は疫学的手法により，それぞれ探求されてきた。しかし，ゲノム科学は「失われた遺伝率」問題により，また疫学は小効果のリスク因子の再現性の危機により，それぞれ限界を迎えつつある。これらの限界を乗り越えると期待されるのが，前向きゲノムコホートとよばれる研究デザインである。前向きゲノムコホートでは定義された集団に対し，ゲノム情報と環境曝露情報を収集し，血液や尿などの生体サンプルを一定の品質管理のもと保管する（バイオバンキング）。また疾患発症情報や，検査値・画像・アンケート結果などの多様で多層的な中間表現型情報を前向きに取得していく。これにより，遺伝子・環境相互作用も含めた解析が可能となり，新たなリスク因子が同定されると期待される。しかしゲノムコホート研究には特有の困難が存在する。一つは p ＞＞ n 問題であり，もう一つは多様で多層的な中間表現型情報から意味のある特徴量を抽出する問題である。これらの問題を解決すると期待されるのが統計的機械学習および深層学習技術である。本稿ではこれらの技術を適用した筆者らの研究例を紹介する。

日本生物学的精神医学会誌 32（2）：81–84, 2021

Key words：prospective genome cohort，biobank，common complex disease，machine learning，artificial intelligence

精神疾患をはじめとした遺伝要因・環境要因が複雑に寄与する有病率の高い疾患は，ありふれた複雑な疾患（common complex diseases）とよばれる。ありふれた複雑な疾患の遺伝要因はゲノム科学的手法により，一方の環境要因は疫学的手法により，それぞれ探求されてきた。ありふれた複雑な疾患に対するゲノム科学の代表的な解析手法は，一塩基多型（single nucleotide polymorphism：SNP）を用いたゲノムワイド関連解析（genome-wide association study：GWAS）である。SNP-GWAS 法においては罹患者集団と対照集団を用意し，罹患者集団に有意に濃縮された遺伝的多型をリスクバリアントとして同定する。もしくは一つの集団において量的形質を目的変数とし，各 SNP 座位のマイナーアリルコピー数（0，1，2）を説明変数として線形回帰を行い，有意な座位を同定する。しかし SNP-GWAS 法がこれまでに同定してきたリスクバリアントは一般に，たとえ有意であっても，その効果量は非常に小さい。そのため同定されたすべてのリスクバリアントの効果量を合算しても，古典的な家系分析で推定された遺伝要因の割合，すなわち遺伝率の大半を説明できないという「失われた遺伝率」問題が生じる[4]。この「失われた遺伝率」問題は長らくゲノム科学研究者を悩ませてきた。レアバリアントの寄与，見逃されてきた構造多型の寄与，遺伝子間相互作用や遺伝子・環境相互作用の寄与から，家系分析で得られた遺伝率推定値が誤っていた可能性などさまざまな説が提出されている[2]が，決着はついていない。一方

Personalized Medicine in Psychiatry using Biobanks

1）東北大学未来型医療創成センター（〒 980-8573　宮城県仙台市青葉区星陵町 2-1 東北メディカル・メガバンク棟 4 階）Jun Takayama：Tohoku University Advanced Research Center for Innovations in Next-Generation Medicine. Tohoku Medical Megabank Building, 4th floor, 2-1 Seiryo-machi, Aoba-ku, Sendai, Miyagi 980-8573, Japan

2）理化学研究所革新知能統合研究センター（〒 103-0027　東京都千代田区日本橋 1-4-1 日本橋一丁目三井ビルディング 15 階）Jun Takayama, Gen Tamiya：RIKEN Center for Advanced Intelligence Project. Nihonbashi 1-chome Mitsui Building, 15th floor, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan

3）東北大学大学院医学系研究科（〒 980-8575　宮城県仙台市青葉区星陵町 2-1）Gen Tamiya：Tohoku University Graduate School of Medicine. 2-1 Seiryo-machi, Aoba-ku, Sendai, Miyagi 980-8575, Japan

【田宮　元　E-mail：gtamiya@med.tohoku.ac.jp】

の疫学研究においても，飲酒や喫煙などのよく知られた大きな効果を持つリスク要因はすでに取り尽くされ，小さな効果のリスク要因に研究の対象が移行している。しかし，これら小効果のリスク因子は研究間での再現性に乏しく，古典的な疫学は危機に瀕していると言われる[6]。このように，ありふれた複雑な疾患に対して，ゲノム科学はその遺伝的側面から，疫学は環境要因の側面からリスク要因の探究を進めてきたが，それぞれに限界を迎えつつある。

これらの問題を乗り越えるためには，遺伝要因・環境要因を同時に，しかも体系的に扱う研究デザインが必要である。その有力な一つが前向きゲノムコホートとよばれる研究デザインである[1]。前向きゲノムコホートでは，10 万人規模の特定の集団に対し，血液・唾液・尿などの生体サンプルを品質管理された方法で取得し保管する（バイオバンキング）。そうしてこれら検体のゲノム情報・環境曝露情報とともに，疾患発症を含むさまざまな表現型情報を体系的に，また経時的に収集する（**図 1**）。前向きゲノムコホートの例としては，英国の UK Biobank（50 万人規模），米国の All of Us（100 万人規模），フィンランドバイオバンク（50 万人規模），わが国の東北メディカル・メガバンク（15 万人規模）などが挙げられる。これら国家規模の前向きゲノムコホートにより，これまで見いだすことのできなかったリスク因子，特に遺伝子・環境相互作用を同定することができると期待される。しかし，前向きゲノムコホートデータの解析には特有の困難が存在する。本稿では前向きゲノムコホート研究における代表的な 2 つの困難を概説し，それらを克服する試みについて紹介する。

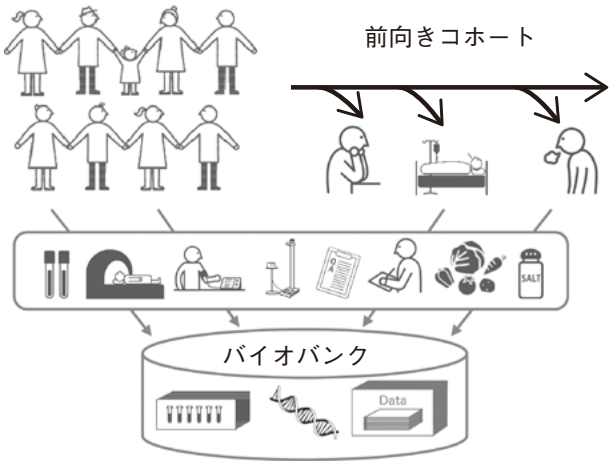

図 1　前向きゲノムコホートとバイオバンク。ベースライン調査において集められたコホートに対し，前向きに追跡調査を行う。また生体サンプルや情報を品質管理された方法でバイオバンクに保存・管理する。

1. p ＞＞ n 問題

1 つ目の問題は p ＞＞ n 問題である。p はパラメータすなわち説明変数の数（次元）を意味し，n は検体数（サンプルサイズ）を意味する。前向きゲノムコホートにおける p はすべての遺伝的多型の座位数と環境曝露の数を合算したものを意味し，その数は数百万～数千万次元に及ぶ。特に次世代シークエンシング法などのゲノム解析手法の飛躍的な進展により，その数はますます大きくなっている。一方の検体数 n は数千から数十万のオーダーである。これらのデータに，単純な重回帰分析の枠組みを適用すると，各検体 i（i ＝ 1, 2, ..., n）の，ある表現型の値 y_i を，p 個の説明変数 $x_1, x_2, ... x_p$ の重み付け和で表すことになる。すなわち $y_i = a_1x_1 + a_2x_2 + ... + a_px_p$ と表される n 本の連立方程式を解くことになるが，p ＞＞ n であるため，解は不定となってしまう。つまり複数の重み付けの組合せが同様に表現型を説明できるため解が 1 つに定まらず，現実には複数あるうちのある 1 つの適合した結果のみを算出してしまい，真に正しいモデルに到達できない。この問題が p ＞＞ n 問題の根本であり，機械学習分野では次元の呪いとして知られてきたものである。この問題を回避するため，SNP-GWAS では各説明変数について単回帰の枠組を適用し，これを説明変数の次元数 p 回繰り返し，多重比較を補正するために Bonferroni 補正を行う。SNP-GWAS ではゲノムワイド有意水準としてしばしば $P < 5 \times 10^{-8}$ が使用されるが，その正当化は慣習的に使われる $P < 0.05$ に 100 万回の多重比較の補正を行うことに基づく。しかし，実際には p 個の説明変数は独立ではなく，それらの間には相関構造が存在するため，この補正は過剰に保守的となり真のリスクバリアントを見逃すことにつながる。統計学分野において，このような説明変数の数がサンプルサイズよりも大きい場合に生じる問題を解決するための手法がこれまでにも研究されてきた。例えばラッソ回帰（least absolute shrinkage and selection operator：Lasso）やその変法である elastic net などの罰則付き回帰法は，バイアスが入ることを一定程度許容することで，一部の説明変数の係数を 0 とし，説明変数を減らすことを実現する手法である。しかしゲノム科学における説明変数は数百万次元と非常に膨大であり，さらに遺伝子間相互作用や遺伝子・環境相互作用を考慮すると 1 兆次元を超える説明変数が生じてしまう。こ

れは統計学分野における研究が想定していたpとnのアンバランスの度合いをはるかに超えるものであり，ゲノム解析に適した新しい統計的手法の開発を必要としていた。

p＞＞n問題を解決するための手法として，筆者らのグループでは，超高次元変数選択法であるsure independence screening（SIS）に基づくゲノム解析手法を開発した。SISにおいては周辺回帰の後に罰則付き回帰を行うことで，最適なモデルを求める。筆者らはSIS法を，超並列計算が可能なGPUを用いて計算できるよう実装し，Wellcome Trust Case-Control ConsortiumやStevens-Johnson症候群のデータに適用し新たなリスクバリアントおよび遺伝子間相互作用を現実的な時間内に同定することに成功した[7, 9]。

p＞＞n問題はリスクバリアントの同定だけでなく，同定したリスクバリアントからの発症リスク予測においても問題となる。発症リスク予測においては，予測に寄与するバリアントを最適な数だけ選ぶ必要がある。筆者らはこれを実現する手法として，smooth-threshold multivariate genetic prediction（STMGP）法を開発した[8]。STMGP法は機械学習手法に基づき，限定的な効果を示すSNP座位に対して適切な重み付けを行いリスク予測に組み入れるものである。シミュレーションを用いた性能評価では，STMGP法はしばしばゲノム予測に使われるgenomic best linear unbiased prediction（GBLUP）法よりも高い性能を示した。また，実際のアルツハイマー病のデータ（Alzheimer's disease neuroimaging initiative：ADNI）や，うつ症状のゲノムリスク予測に対してもより高い予測性能を示した[5, 8]。

2. 多様で多層的な中間表現型からの要約問題

前向きゲノムコホート研究においては多様な研究計画を実行できるように，さまざまな表現型情報を取得する。例えば身長・体重・生年月のような基礎的な情報，血液生化学検査の検査値から，脳MRI画像やアンケート調査に至るまで多岐にわたる情報が収集される。これら表現型のデータは分子レベル，細胞レベル，臓器レベル，個体レベルとさまざまな階層から収集され，また複数の階層にまたがるものも多い。これら個別の値を目的変数とした解析ももちろん可能であるが，生体のより本質的な特徴を捉えるためには，個別の値だけではなくデータ全体をみる必要がある。特に精神科領域の疾患のような場合にはその重要性は高い。これは医師が臨床の現場でさまざまな検査値，画像解析結果や問診結果を解釈しながら，診断仮説を組み立てることに相当する。遺伝統計学の文脈においては，潜在変数として仮定される易罹病性（liability）変数を，個別の表現型変数から合成し，特定の域値を超えるかどうかをもって疾患の有無を判定することに相当する。このようなことが可能となれば，前向きゲノムコホートが収集した表現型データを真に有効活用することが可能となる。しかし多様で多層的な中間表現型データを有効に活用するための手法は確立していない。

統計学の分野では複数の変数から内在する構造を見いだし，より低次元の変数を合成する研究が行われてきた。代表的な手法としては，主成分分析（principal component analysis：PCA）が挙げられる。例えば筆者らは，PCAをイネの構造に関する形質値に適用して得た主成分に対するGWASを行い遺伝子の同定に成功している[11]。しかし，PCAは線形の手法であり，より複雑な構造を持った表現型データ，特に表現型間に非線形の構造がある場合にはより発展的な手法が適しているかもしれない。オートエンコーダは，深層学習の枠組みを用いた次元削減手法であり，深層学習研究ブームのきっかけとなったものである[3]。オートエンコーダでは多次元の入力値をニューラルネットワークに入力し，同数の次元で出力する。この入力層と出力層をつなぐネットワークの中間層のなかに，ごく少数のニューロンからなる層が存在する。これはPCAにおける主成分に相当するとみなすことができ，データに内在する構造を捉えた低次元の表現として捉えることが可能である。例えば筆者らの研究ではオートエンコーダを用いて，前立腺がんの病理画像から新しい組織学的特徴を抽出することに成功している[10]。これらの成功は前向きゲノムコホートにおける中間表現型からの適切な特徴量の抽出に深層学習の枠組みが有効である可能性を示唆している。さらに今後は，画像と検査値・アンケート結果などのように大きく種類の異なるデータ（マルチモーダルデータ）からの適切な特徴量の抽出手法の開発が重要となるだろう。

本論文に記載した筆者らの研究に関してすべて倫理的配慮を行っている。開示すべき利益相反は存在しない。

文　献

1）Collins FS（2004）The case for a US prospective cohort study of genes and environment. Nature, 429（6990）：475-477.

2）Eichler EE, Flint J, Gibson G, et al（2010）Missing heritability and strategies for finding the underlying causes of comples disease. Nat Rev Genet, 11：446-450.
3）Hinton GE and Salakhutdinov RR（2006）Reducing the dimensionality of data with neual networks. Science, 313：504-507.
4）Maher B（2008）Personal genomes：The case of the missing heritability. Nature, 456：18-21.
5）Takahashi Y, Ueki M, Tamiya G, et al（2020）Machine learing for effectively avoiding overfitting is a crucial strategy for the genetic prediction of polygenic psychiatric phenotypes. Trans Transl Psychiatry, 10：294.
6）Taubes G（1995）Epidemiology faces its limits. Science, 269（5221）：164-169.
7）Ueki M and Tamiya G（2012）Ultrahigh-dimensional variable selection method for whole-genome gene-gene interaction analysis. BMC Bioinform, 13：72.
8）Ueki M and Tamiya G（2016）Smooth-threshold multivariate genetic prediction with unbiased model selection. Genet Epidemiol, 40：233-243.
9）Ueta M, Katsushi T, Sotozono C, et al（2012）Epistatic interaction between Toll-like receptor 3（TLR3）and prostaglandin E receptor 3（PTGER3）genes. J Allergy Clin Immunol, 129：1413-1416.
10）Yamamoto Y, Tsuzuki T, Akatsuka J, et al（2019）Automated acquisition of explainable knowledge from unannotated histopathology images. Nat Commun, 10：5642.
11）Yano K, Morinaka Y, Wang F, et al（2019）GWAS with principal component analysis identifies a gene comprehensively controlling rice architecture. Proc Natl Acad Sci USA, 116：21262-21267.

■ ABSTRACT

Personalized Medicine in Psychiatry using Biobanks

Jun Takayama[1,2], Gen Tamiya[2,3]

1) Tohoku University Advanced Research Center for Innovations in Next-Generation Medicine
2) RIKEN Center for Advanced Intelligence Project
3) Tohoku University Graduate School of Medicine

For complex diseases, including psychiatric disorders, genetic and environmental risk factors were investigated using genomic and epidemiological methods. However, genomics is faced by the so-called “missing heritability” problem, and epidemiology is limited by the reproducibility crisis for small effect risk factors. A research design called prospective genome cohort is expected to overcome these limitations. Prospective genomic cohorts collect genomic and environmental exposure information for a defined population and store biological samples such as blood and urine under a controlled quality. Besides, genome cohorts prospectively obtain disease onset information and various endophenotypes such as laboratory tests, brain imaging, and questionnaire survey. With the prospective genome cohort, new risk factors, including gene-environment interactions, will be identified. However, genome cohort studies have unique difficulties. One is the p >> n problem, and the other is the problem in extracting meaningful features from diverse and multi-layered endophenotypic information. Here, we will present examples of statistical machine learning and deep learning techniques that are expected to address these problems.

（Japanese Journal of Biological Psychiatry 32（2）：81-84, 2021）