Tổng quan ứng dụng học máy trong dự đoán nguy cơ đa di truyền hướng tới y học cá thể hóa

Trịnh Thị Xuân[1], Tạ Văn Nhân[2], Hoàng Đỗ Thanh Tùng[3], Trương Nam Hải[4], Trần Đăng Hưng[5]
[1] Khoa Công nghệ thông tin, Đại học Mở Hà Nội
[2] Công ty TNHH LOBI Việt Nam, Hà Nội
[3] Phòng Nghiên cứu hệ thống và quản lý, Viện Công nghệ Thông Tin, VAST
[4] Phòng Kỹ thuật di truyền, Viện Công nghệ Sinh học, VAST
[5] Khoa Công nghệ thông tin, Đại học Sư phạm Hà Nội

Tác giả liên hệ: Trịnh Thị Xuân, Email: trinhxuan@gmail.com
Ngày nhận bài: 06/01/2022, ngày sửa chữa: 21/04/2022, ngày duyệt đăng: 10/05/2022
Định danh DOI: 10.32913/mic-ict-research-vn.v2022.n1.1027

Tóm tắt: Trong thời gian gần đây, Điểm nguy cơ đa di truyền (Polygenic Risk Score - PRS) được xem như một công cụ tiềm năng cho y học chính xác dựa trên các biến dị di truyền phổ biến có đóng góp từ nhỏ tới vừa đối với nguy cơ mắc bệnh di truyền, nhưng tổng gộp các biến dị này lại có thể nâng cao giá trị dự đoán bệnh trong quần thể. Đã có nhiều phương pháp học máy được đưa ra nhằm cải tiến khả năng dự đoán của PRS cũng như những nỗ lực để đưa PRS vào ứng dụng trong lâm sàng. Mặc dù vậy, việc lựa chọn phương pháp một cách hệ thống và những ứng dụng của PRS vẫn chưa thực sự rõ ràng. Vì vậy, trong bài báo tổng quan này, chúng tôi cung cấp một cái nhìn tổng quan về điểm nguy cơ đa di truyền và các nghiên cứu cải tiến sử dụng học máy nhằm nâng cao khả năng áp dụng trong lâm sàng của PRS.

Từ khóa: *Bệnh phổ biến, điểm nguy cơ đa di truyền, GWAS, SNP, mảng SNP, học máy.*

Title: **An Overview of Machine Learning Applications in Polygenic Risk Prediction Towards Personalized Medicine**

Abstract: In recent times, the Polygenic Risk Score (PRS) has been considered as a potential tool for precision medicine based on common genetic variants with small to moderate contributions to the genetic disease risk, but the aggregation of these variations can enhance the predictive value of disease in the population. Many machine learning methods have been proposed to improve the predictive ability of PRS as well as efforts to bring PRS into clinical application. However, the systematic selection of methods and applications of PRS are still not really clear. Therefore, in this review, we provide an overview of polygenic risk scores and innovative studies using machine learning to improve the clinical applicability of PRS.

Keywords: *Common diseases, polygenic risk scores, GWAS, SNP, SNP arrays, machine learning.*

I. GIỚI THIỆU

Với sự phát triển của công nghệ sinh học, các nhà khoa học có thể dựa vào dữ liệu DNA để dự đoán nguy cơ mắc bệnh ở người. Ngoại trừ các đột biến xôma, dữ liệu DNA không thay đổi trong suốt thời gian sống của chúng ta. Vì vậy, nguy cơ mắc bệnh liên quan đến gen của một người có thể xuất hiện ngay từ khi người đó sinh ra. Điều đó cho thấy việc xác định nguy cơ mắc bệnh về gen có ý nghĩa rất lớn trong y tế dự phòng. Chẳng hạn, một nghiên cứu vào năm 2017 ước lượng rằng có khoảng 72% phụ nữ thừa hưởng đột biến gen *BRCA1* và khoảng 69% phụ nữ thừa hưởng đột biến gen *BRCA2* sẽ phát triển ung thư trước tuổi 80 [1]. Việc phát hiện một bệnh nhân mang các biến dị có hại sẽ hỗ trợ bác sỹ đưa ra lời khuyên giúp thay đổi lối sống theo hướng tích cực hoặc đưa ra các can thiệp phòng ngừa tùy theo mức độ nguy cơ.

Với các bệnh do một gen quy định, việc ước lượng nguy cơ mắc bệnh của một người có thể chỉ là tìm kiếm các biến dị có hại trên một gen nào đó. Các nghiên cứu phân tích liên kết di truyền (Genetic Linkage Analysis) đã được phát triển từ lâu để xác định vị trí của gen bệnh dựa trên mối liên kết của nó với các vị trí đánh dấu di truyền trên

nhiễm sắc thể. Phương pháp này đã rất thành công khi tìm ra các đột biến của một số bệnh đơn gen như múa giật [2, 3] hay ung thư vú [4]. Tuy nhiên, phân tích liên kết chưa cho thấy sự hiệu quả đối với các bệnh đa di truyền và phổ biến. Từ năm 2005, các nghiên cứu tương quan toàn hệ gen (Genome-Wide Association Studies - GWAS) đã bắt đầu tìm kiếm các đa hình đơn nucleotit (Single-Nucleotide Polymorphism, SNP) phổ biến (tần số alen phụ, minor allele frequency, MAF $\geq$ 1%) [5–7]. GWAS ngày càng được tạo điều kiện thuận lợi nhờ sự phát triển của các mảng SNP (SNP array) với chi phí tương đối thấp, các mảng này có thể chứa từ 200.000 đến 2.000.000 SNP [8]. Ngoài ra, quá trình phân tích tổng hợp các nghiên cứu tương quan toàn hệ gen riêng lẻ cũng góp phần tạo ra các bộ dữ liệu GWAS ngày càng lớn hơn. Các SNP bị thiếu trên mảng SNP sẽ được bổ sung thông qua quá trình suy diễn thống kê từ các SNP đã được quan sát và các kiểu gen đơn bội (Haplotype) của dữ liệu tham chiếu Chính nhờ quá trình này mà ta có được các tập dữ liệu GWAS lớn như UK Biobank: 96 triệu biến dị cho gần 500 nghìn cá thể mà ban đầu chỉ có khoảng 800 nghìn SNP được xác định bởi kiểu gen [9]. Mặc dù vậy, GWAS vẫn tỏ ra hạn chế trong dự đoán các bệnh đa di truyền, cho dù sử dụng các SNP có tương quan cao với bệnh thì dự đoán cũng không thực sự chính xác [10, 11]. Để nâng cao hiệu suất dự đoán, các nghiên cứu tập trung vào việc lựa chọn tập SNP đặc trưng, thực sự ảnh hưởng đến bệnh. Các phương pháp truyền thống thường dựa vào đặc điểm sinh học và thống kê, điển hình như loại bỏ các SNP do mất cân bằng liên kết hoặc do di truyền cùng nhau mà không ảnh hưởng đến bệnh, giữ lại các SNP có tương quan cao với bệnh. Với một bệnh cụ thể, mặc dù mỗi SNP đặc trưng chỉ ảnh hưởng một phần nhỏ tới bệnh, nhưng kết hợp chúng với nhau có thể giải thích một phần đáng kể tỷ lệ mắc bệnh trong quần thể [8, 12–14].

Để lựa chọn tập SNP đặc trưng, ngoài các phương pháp dựa vào đặc điểm sinh học, các phương pháp áp dụng học máy cũng đang nở rộ trong thời gian gần đây. Cụ thể, ta có thể đưa vấn đề về bài toán lựa chọn đặc trưng [15–17], đặc biệt là các phương pháp lựa chọn đặc trưng áp dụng cho dữ liệu chiều cao có số đặc trưng lớn hơn nhiều số mẫu. Đã có nhiều nhóm phương pháp khác nhau được sử dụng như lọc bằng cách đặt ngưỡng [18], loại bỏ đặc trưng ngay trong quá trình học [19–21], cũng như tính đến các ảnh hưởng phi tuyến [22]. Các phương pháp có thể kết hợp một cách khéo léo trong mô hình dự đoán nguy cơ đa di truyền (Polygenic Risk Scores - PRS) [23] để dự đoán nguy cơ mắc bệnh trong một nhóm thuần tập đích sử dụng thống kê tóm tắt GWAS của một nhóm thuần tập cơ sở, độc lập với nhóm thuần tập đích [24, 25]. Sự đa dạng trong các phương pháp góp phần giúp các nhà khoa học có một số nghiên cứu quy mô trong những năm gần đây với nhiều hơn các bệnh di truyền phức tạp, cùng với đó là các tập dữ liệu ngày càng lớn. Với hiệu suất mô hình dự đoán ngày càng được cải thiện, điểm nguy cơ đa di truyền đang góp phần vào nỗ lực phân tầng nguy cơ di truyền và có triển vọng áp dụng rộng rãi trong lâm sàng.

Trong các phần tiếp theo của bài báo chúng tôi sẽ trình bày về các kiến thức cơ sở trong Phần II, về tiền xử lý dữ liệu trong Phần III. Các cải tiến sử dụng học máy để nâng cao khả năng áp dụng trong lâm sàng của PRS được trình bày ở Phần IV. Phần V đề cập đến các xu hướng cải tiến hiệu năng cho mô hình dự đoán nguy cơ đa di truyền trong tương lai. Cuối cùng, chúng tôi kết luận bài báo ở Phần VI.

II. KIẾN THỨC CƠ SỞ

1. Nghiên cứu tương quan toàn hệ gen (GWAS)

Nghiên cứu tương quan toàn hệ gen (GWAS) là một phương pháp sàng lọc nhanh các chỉ thị phân tử trên toàn bộ DNA hoặc bộ gen của nhiều người, mục đích để tìm các biến dị di truyền liên quan đến một căn bệnh cụ thể hoặc một tính trạng mà ta quan tâm. Để thực hiện một nghiên cứu tương quan toàn hệ gen, các nhà nghiên cứu sử dụng hai nhóm người: nhóm thuần tập ca bệnh (case) và thuần tập đối chứng (control). DNA thu được bằng cách tách triết mẫu máu hoặc tế bào niêm mạc miệng của những người tham gia được đặt trên các mảng SNP và được quét trên các máy tự động. Các máy này nhanh chóng khảo sát bộ gen của mỗi người để tìm ra các SNP. Nếu SNP nào được tìm thấy ở những người mắc bệnh với tần suất cao hơn so với những người không mắc bệnh, thì SNP đó được cho là có mối tương quan đến bệnh. Các SNP này có thể đóng vai trò là những điểm đánh dấu di truyền (makers) chỉ dẫn đến khu vực chứa gen bệnh [1].

Do phải tuân theo quy tắc bảo mật của từng dự án nghiên cứu và các chính sách bảo mật thông tin cá nhân mà việc truy cập vào dữ liệu GWAS có hai mức độ:

- Mức độ thống kê tóm tắt: bao gồm các giá trị đại diện cho nhiều điểm dữ liệu, chẳng hạn như mức độ ảnh hưởng và P-value cho sự tương quan giữa mỗi SNP với một kiểu hình mà ta quan tâm. Chú ý rằng, mức độ ảnh hưởng được biểu diễn bằng tỷ lệ chênh lệch (Odds Ratio, viết tắt là OR) với tính trạng rời rạc, và được biểu diễn bằng chỉ số β với tính trạng liên tục. Ngoài ra, dữ liệu còn bao gồm các alen, tên SNP và vị trí của chúng trên nhiễm sắc thể.
- Mức độ cá thể: bao gồm định danh của cá thể, bố, mẹ, và phả hệ của cá thể. Hơn nữa, dữ liệu cũng cung cấp các thông tin về giới tính, kiểu hình, các alen, vị

[1] https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet

trí của các SNP trên nhiễm sắc thể, khoảng cách di truyền cũng như các hiệp biến. Dữ liệu GWAS ở mức độ cá thể thường được lưu dưới dạng các tệp định dạng PLINK [26, 27].

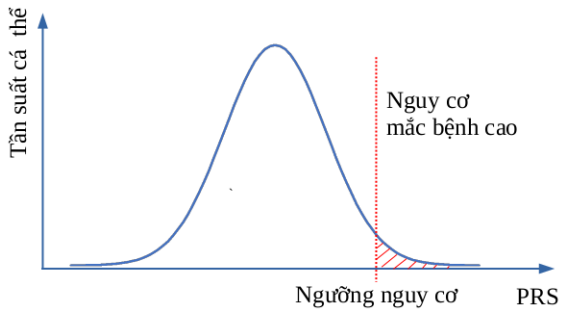
Dữ liệu thống kê tóm tắt GWAS có thể được tra cứu dễ dàng với nhiều tính trạng khác nhau trong Danh mục NHGRI-EBI về các nghiên cứu tương quan toàn hệ gen [28] [2]. Ngược lại, các dữ liệu GWAS ở mức độ cá thể cần được cấp phép để truy cập như tài nguyên trên dbGaP [29] hay UK Biobank [9].

2. Tổng quan về điểm nguy cơ đa di truyền (PRS)

Điểm nguy cơ đa di truyền (Polygenic Risk Score, PRS) được tính bằng tổng điểm có trọng số của các alen nguy cơ với trọng số dựa trên các mức độ ảnh hưởng từ GWAS [24]. Công thức mặc định để tính PRS trong PLINK [26] là:

$$PRS_j = \frac{\sum_i^N S_i.G_{i,j}}{P.M_j}$$

trong đó mức độ ảnh hưởng của SNP thứ i là S_i; số các alen ảnh hưởng của SNP thứ i được quan sát trong mẫu j là $G_{i,j}$; đơn bội của mẫu là P (thường là 2 cho người); tổng số SNP của mẫu j là N; tổng số SNP không thiếu được quan sát ở mẫu j là M_j. Nếu mẫu j có một kiểu gen thiếu SNP thứ i thì tần số alen phụ của quần thể được nhân với đơn bội ($MAF_i.P$) được sử dụng thay thế $G_{i,j}$.



Hình 1. **Biểu đồ phân phối của điểm nguy cơ đa di truyền**. Những cá thể có PRS nằm trong ngưỡng nguy cơ ở phía đuôi phải của phân phối có nguy cơ mắc bệnh đa di truyền cao.

Điểm nguy cơ đa di truyền (PRS) đang được coi là độ đo tương đối để mọi người có thể tìm hiểu về nguy cơ phát triển bệnh của họ. Ta có thể biểu diễn các giá trị PRS dưới dạng biểu đồ phân phối tần suất mà một cá thể có nguy cơ mắc bệnh cao sẽ nằm trong ngưỡng nguy cơ ở đuôi phải của phân phối (Xem hình 1). Tuy nhiên, độ chính xác của các dự đoán nguy cơ mắc bệnh đối với các nhóm người là khác nhau do nó phụ thuộc vào sự đầy đủ của dữ liệu GWAS. Tính đến năm 2018, phần lớn các nghiên cứu tương quan toàn hệ gen đã được thực hiện với tỷ lệ số lượng cá thể dựa trên sắc tộc bao gồm 78% người châu Âu, 10% người châu Á, 2% người châu Phi, 1% người gốc Tây Ban Nha và tất cả các sắc tộc khác chỉ chiếm nhỏ hơn 1% GWAS [30]. Ngoài ra, PRS cũng chỉ ra đóng góp của các yếu tố đa di truyền đối với một kiểu hình mà dữ liệu GWAS không phát hiện được. Vào năm 2009, một GWAS cho bệnh tâm thần phân liệt do Purcell và các đồng nghiệp thực hiện đã tìm thấy chỉ một SNP tương quan với kiểu hình, mặc dù bệnh này được biết đến như là một bệnh có tính di truyền cao. Tuy nhiên, bằng cách xây dựng một PRS sử dụng các kết quả GWAS và kiểm tra điểm nguy cơ đa di truyền trong tập dữ liệu độc lập khác, Purcell đã chứng minh rằng có một đóng góp đa di truyền tới bệnh tâm thần phân liệt [31]. Do đó, phân tích đa di truyền cho thấy dữ liệu GWAS ở giai đoạn đầu là chưa đủ mạnh, mà chúng ta cần có cỡ mẫu lớn hơn [32]. Mục đích khác của việc sử dụng PRS là để kiểm tra mối tương quan giữa các SNP với một kiểu hình khác với kiểu hình được sử dụng trong GWAS. Kỹ thuật này cho phép các nhà nghiên cứu chứng minh rằng có sự đóng góp đa di truyền chung giữa hai tính trạng. Cụ thể, Purcell đã chứng minh rằng có sự đóng góp di truyền chung giữa bệnh tâm thần phân liệt và rối loạn lưỡng cực (xem Hình 2).

Nếu như nghiên cứu của Purcell đã chỉ ra một số đặc điểm cơ bản là nền móng của PRS thì gần đây một số nghiên cứu lớn đã cải thiện một cách đáng kể độ tin cậy của kết quả dự đoán. Năm 2018, Amit V. Khera và các đồng nghiệp đã đã sử dụng dữ liệu di truyền toàn bộ hệ gen và các phương pháp suy diễn thống kê để đánh giá hàng triệu biến dị di truyền phổ biến liên quan đến 5 bệnh phổ biến: động mạch vành, rung nhĩ, tiểu đường loại 2, viêm ruột, và ung thư vú. Đối với mỗi bệnh, họ đã áp dụng một thuật toán tính điểm nguy cơ đa di truyền từ tất cả các biến dị để phản ánh tính nhạy cảm di truyền của một người đối với những bệnh này [33]. Năm 2019, một nghiên cứu với 272 tác giả và tổ chức sử dụng dữ liệu GWAS về ung thư vú được coi là lớn nhất với tập hợp từ 69 nghiên cứu gồm 94,075 mẫu bệnh và 75,017 mẫu đối chứng. Kết quả đã tìm ra được 313 SNP cho tính toán PRS tốt nhất với AUC đạt 63%, độ tin cậy 95%. Ngoài ra, nhóm tác giả đã phân tầng nguy cơ mắc bệnh dựa trên cả lịch sử gia đình và các yếu tố nguy cơ khác [34]. Một số nghiên cứu đột phá trong lịch sử phát triển của phương pháp tính điểm nguy cơ đa di truyền ở trên được tóm tắt trong Bảng I.

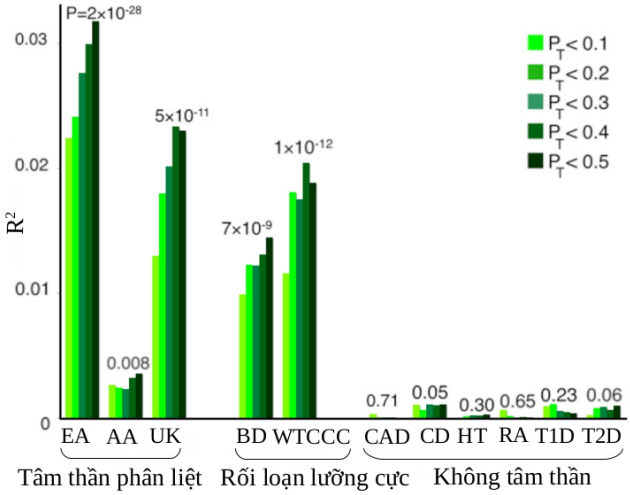
Để xây dựng mô hình tính điểm nguy cơ đa di truyền ta cần hai loại dữ liệu:

- Dữ liệu cơ sở (base data): bao gồm các thống kê tóm tắt của GWAS (ví dụ, β, OR, P-values) của tương quan kiểu gen-kiểu hình tại một biến dị di truyền (SNP).
- Dữ liệu đích (target data): kiểu gene, kiểu hình của

[2] https://www.ebi.ac.uk/gwas/

Bảng I
CÁC NGHIÊN CỨU ĐỘT PHÁ TRONG TÍNH TOÁN PRS.

STT	Tác giả	Năm	Nội dung
1	S. M. Purcell	2009	Xác định ảnh hưởng đa di truyền chung giữa bệnh tâm thần phân liệt và rối loạn lưỡng cực [31].
2	A. V. Khera	2018	Đánh giá hàng triệu biến dị phổ biến liên quan đến năm bệnh đa di truyền phổ biến [33].
3	N. Mavaddat	2019	Phân tầng nguy cơ ung thư vú dựa trên dữ liệu lớn của mẫu bệnh và mẫu đối chứng [34].



Hình 2. **Các thành phần đa di truyền với các mẫu độc lập của bệnh tâm thần phân liệt và rối loạn lưỡng cực**. Có sự đóng góp đa di truyền chung giữa bệnh tâm thần phân liệt và rối loạn lưỡng cực. Hình ảnh được tham khảo từ nghiên cứu của Hiệp hội Tâm thần Quốc tế [31].

các cá thể.

Một bệnh phát sinh ngoài nguyên nhân do di truyền thì còn có các nguyên nhân khác như môi trường, lối sống... Do đó bệnh không chỉ được dự đoán hoàn toàn nhờ các yếu tố di truyền, mà các yếu tố này chỉ chiếm một tỷ lệ nhất định h^2 (cũng được gọi là hệ số di truyền SNP: h^2_{SNP}, xem định nghĩa ở hộp 1). Vì việc xác định tập biến dị ảnh hưởng hoặc ước lượng mức độ ảnh hưởng có thể chưa chính xác, cũng như có sự khác biệt giữa các mẫu cơ sở và các mẫu đích mà tỷ lệ dự đoán của PRS nhỏ hơn h^2_{SNP}. Để đảm bảo mô hình PRS có được hiệu suất dự đoán cao ta cần kiểm tra kỹ lưỡng và tiền xử lý dữ liệu theo các tiêu chí sau:

- Các mẫu trong dữ liệu cơ sở và dữ liệu đích cần độc lập với nhau.
- Dữ liệu cơ sở và dữ liệu đích có các vị trí SNP được xây dựng từ cùng một hệ gen tham chiếu.

Hộp 1 | Một số khái niệm cơ bản

Alen nguy cơ (Risk Allele): alen của SNP làm tăng nguy cơ mắc bệnh. Một alen ảnh hưởng chỉ là alen có tương quan với bệnh và có thể tăng hoặc giảm nguy cơ.
Tần số alen (Allele Frequency): tỷ lệ các nhiễm sắc thể mang alen đó trong một quần thể.
Tần số alen phụ (MAF): tần số mà alen phổ biến thứ hai xuất hiện trong một quần thể nhất định.
Biến dị nhân quả (Causal Variant): biến dị tại một vị trí xác định trong nhiễm sắc thể có tương quan mạnh với kiểu hình, các biến dị này thực sự có ảnh hưởng sinh học đến kiểu hình [35].
Mức độ ảnh hưởng (Effect Size): mức độ tương quan giữa tần số alen của biến dị với một bệnh nào đó.
Hệ số di truyền (Heritability): mức độ giải thích của các yếu tố di truyền với kiểu hình trong một quần thể (ký hiệu: h^2 hoặc h^2_{SNP}) [36].
Mất cân bằng liên kết (Linkage Disequilibrium, LD): thước đo sự kết hợp không ngẫu nhiên giữa các alen ở các vị trí khác nhau trên cùng một nhiễm sắc thể trong một quần thể nhất định. Các SNP có trong LD khi tần số tương quan của các alen của chúng cao hơn mong đợi [27].
Phân tầng quần thể (Population Stratification): sự hiện diện của nhiều quần thể con (ví dụ: các cá nhân có nguồn gốc dân tộc khác nhau) trong một nghiên cứu. Sự phân tầng quần thể có thể dẫn đến các tương quan dương tính giả.
ROC (Receiver Operating Characteristic Curve): đường cong đặc tính biểu diễn tương quan giữa dương tính giả và dương tính thật với các ngưỡng nào đó [37, 38].
AUC (Area Under the Curve): diện tích dưới đường ROC, giá trị AUC nằm trong khoảng (0, 1), giá trị này càng lớn thì hiệu quả của mô hình càng cao [39].
Precision: tỉ lệ số điểm dương tính thật trên số điểm được phân loại là dương tính (dương tính thật + dương tính giả).
Recall: tỉ lệ số điểm dương tính thật trên số điểm thực sự là dương tính (dương tính thật + âm tính giả).
Overfitting: hiện tượng mô hình tìm được quá khớp với dữ liệu huấn luyện dẫn đến độ chính xác của dự đoán không còn tốt trên dữ liệu kiểm thử. Vì tần số alen có thể khác nhau giữa các quần thể con nên sự phân tầng quần thể có thể dẫn đến các tương quan dương tính giả.

- Thực hiện đầy đủ quá trình kiểm soát chất lượng cho dữ liệu cơ sở và dữ liệu đích.

Ngoài ra, phân tầng quần thể, yếu tố gây nhiễu trong GWAS, lại có thể được sử dụng để tính toán PRS tốt hơn.

3. Ứng dụng của PRS trong công nghiệp

PRS có thể được ứng dụng để dự đoán nguy cơ mắc bệnh trong công nghiệp từ mẫu nước bọt hoặc mẫu máu sử dụng công nghệ định kiểu gen (Genotyping) không tốn kém. Mặc dù vậy, PRS không bao giờ có thể dự đoán chắc chắn về tình trạng của các bệnh phổ biến phức tạp vì các yếu tố di truyền chỉ đóng góp một phần nguy cơ và PRS cũng chỉ nắm bắt được một phần các đóng góp di truyền đó.Tuy nhiên, cũng giống như y học lâm sàng sử dụng vô số các biện pháp dự đoán khác, PRS đóng một vai trò đáng kể như là một phần của các thuật toán dự đoán đa biến [40].

Bảng II
10 BỆNH HOẶC TÌNH TRẠNG Y TẾ ĐƯỢC FDA THÔNG QUA CHO CÁC XÉT NGHIỆM VỀ NGUY CƠ MẮC BỆNH DI TRUYỀN CỦA 23ANDME.

Bệnh	Mô tả
Parkinson	Rối loạn hệ thống thần kinh ảnh hưởng đến chuyển động
Alzheimer khởi phát muộn	Chứng rối loạn não tiến triển phá hủy trí nhớ và kỹ năng tư duy
Celiac	Rối loạn dẫn đến không thể tiêu hóa gluten
Thiếu Alpha-1 antitrypsin	Rối loạn làm tăng nguy cơ mắc bệnh phổi và gan
Loạn trương lực cơ nguyên phát khởi phát sớm	Rối loạn vận động liên quan đến các cơn co thắt cơ không tự chủ và các cử động mất kiểm soát khác
Thiếu yếu tố XI	Rối loạn đông máu
Gaucher loại 1	Rối loạn cơ quan và mô
Thiếu hụt Glucose-6-Phosphate Dehydrogenase	Còn được gọi là G6PD, một tình trạng hồng cầu
Huyết sắc tố di truyền	Rối loạn quá tải sắt
Máu khó đông di truyền	Rối loạn đông máu

Năm 2017, Cơ quan Quản lý Thực phẩm và Dược phẩm Hoa Kỳ (U.S. Food and Drug Administration - FDA) đã cho phép tiếp thị các xét nghiệm nguy cơ sức khỏe di truyền (Genetic Health Risk, GHR) của dịch vụ hệ gen cá nhân 23andMe đối với 10 bệnh hoặc tình trạng y tế (Xem Bảng II). Đây là các xét nghiệm đầu tiên được FDA cho phép cung cấp thông tin về khuynh hướng di truyền của một cá nhân đối với một số bệnh hoặc tình trạng y tế nhất định [3]. Các xét nghiệm GHR của 23andMe hoạt động bằng cách phân lập DNA từ mẫu nước bọt, sau đó được kiểm thử với hơn 500,000 biến dị di truyền. Sự hiện diện hoặc vắng mặt của một số biến dị này có liên quan đến việc tăng nguy cơ phát triển của bệnh.

Tại hội nghị South by Southwest (SXSW) 2019, 23andMe đã thông báo họ sẽ cung cấp phân tích nguy cơ dựa trên di truyền với bệnh tiểu đường loại 2 cho hàng triệu khách hàng. Thông tin di truyền từ các cá nhân với dữ liệu liên quan đến sức khỏe cũng hỗ trợ việc xây dựng các mô hình thống kê cho tính toán PRS và dự đoán các tính trạng và các tình trạng y tế khác nhau từ một DNA cá nhân. Đặc biệt trong báo cáo về bệnh tiểu đường loại 2 vào năm 2022, 1.1 triệu khách hàng của 23andMe đã đồng ý tham gia nghiên cứu giúp tạo ra cơ sở dữ liệu kiểu gen-kiểu hình lớn với 11,999 biến dị di truyền. Nhờ đó, các nhà nghiên cứu của 23andMe đã nâng cao được độ chính xác trong tính toán PRS với hơn 1000 biến dị liên quan đến bệnh [4].

III. TIỀN XỬ LÝ DỮ LIỆU

1. Kiểm soát chất lượng (Quality Control, QC)

Độ chính xác dự đoán của PRS phụ thuộc lớn vào chất lượng của dữ liệu cơ sở và dữ liệu đích. Do đó, đã có nhiều nghiên cứu cho quá trình kiểm soát chất lượng (Quality Control, QC). Cả hai tập dữ liệu thường được tiến hành QC với các tiêu chuẩn QC chung của GWAS [27, 41, 42]. Vào năm 2020, Shing Wan Choi và các đồng nghiệp đã chia ra ba cấp QC liên quan đến từng loại dữ liệu [43]. Dựa vào các bước QC từ những nghiên cứu trước đây, chúng tôi đã sắp xếp lại để đưa ra một quy trình kiểm soát chất lượng rõ ràng và đầy đủ (Xem Hình 3).

- **QC chỉ liên quan đến dữ liệu cơ sở**: gồm kiểm tra hệ số di truyền và xác định các alen ảnh hưởng.
- **QC chỉ liên quan đến dữ liệu đích**: cỡ mẫu được khuyến nghị lớn hơn hoặc bằng 100 [44] và thận trọng với những phân tích sử dụng dữ liệu cơ sở có h^2_{SNP} thấp và cỡ mẫu dữ liệu đích nhỏ. Mặt khác, cần kiểm soát chất lượng trên một số tiêu chí như: tỉ lệ kiểu gen (Genotyping Rate > 0.99), tỷ lệ mẫu thiếu (Sample Missingless < 0.02), cân bằng Hardy-Weinberg (P > 10^{-6}), loại bỏ những cá thể có hệ số F (đại diện cho tỷ lệ dị hợp tử) lớn hoặc nhỏ hơn 3 lần độ lệch chuẩn so với trung bình, loại bỏ các SNP tương quan với $r^2 > 0.25$.
- **QC tiêu chuẩn**: áp dụng cho cả dữ liệu cơ sở và dữ liệu đích. Do dữ liệu lớn, ta cần đảm bảo dữ liệu không bị lỗi trong quá trình tải hoặc chuyển tệp. Các dữ liệu cần được QC theo các tiêu chí chung cho dữ liệu GWAS như: tần số alen phụ (MAF > 1%, hoặc MAF > 5% nếu số lượng mẫu đích nhỏ hơn 1000), điểm thông tin (Info Score) bổ sung (Imputation) lớn

[3] https://www.fda.gov/news-events/press-announcements/fda-allows-marketing-first-direct-consumer-tests-provide-genetic-risk-information-certain-conditions

[4] https://blog.23andme.com/23andme-research/screening-for-t2d/

hơn 0.8. Ngoài ra, ta cần loại bỏ các cá thể có sự khác biệt giới tính, xử lý các SNP không khớp, loại bỏ các SNP trùng lặp, loại bỏ các mẫu trùng lặp hoặc các mẫu có quan hệ họ hàng.

Dữ liệu cơ sở và dữ liệu đích sau khi được QC riêng sẽ tham gia quá trình QC tiêu chuẩn để tạo ra dữ liệu cuối cùng phục vụ cho tính toán điểm nguy cơ đa di truyền.

2. Kiểm soát mất cân bằng liên kết (Linkage Disequilibrium, LD)

Sau quá trình QC trên dữ liệu cơ sở và dữ liệu đích và định dạng các tập dữ liệu một cách phù hợp, ta cần lựa chọn một tập hợp các SNP ảnh hưởng trên toàn hệ gen để phục vụ tính toán PRS. Công việc này khá phức tạp do mối tương quan giữa các SNP được sinh ra từ sự mất cân bằng liên kết. Ta có thể lựa chọn tập hợp các SNP ảnh hưởng thông qua:

- Tính toán mất cân bằng liên kết (Linkage Disequilibrium, LD).
- Thu hẹp mức độ ảnh hưởng được ước lượng bởi GWAS (Shrinkage).

Phương pháp truyền thống thường được sử dụng là "Nhóm và Đặt ngưỡng" ("Clumping and Thresholding", hay còn gọi là "C+T"). Các kiểm định mối tương quan trong GWAS được thực hiện từng SNP một, trong khi cấu trúc tương quan trên toàn bộ hệ gen rất phức tạp, nên việc ước lượng các tác động di truyền độc lập là một vấn đề thách thức. Các SNP có thể được "Nhóm" ("Clumping", C) bởi công cụ PLINK để chọn ra các SNP có mối tương quan thấp với nhau. Trước tiên, Clumping chọn ra một SNP đặc trưng được gọi là SNP chỉ mục (SNP index) và tính toán mối tương quan giữa SNP này với các SNP gần đó (trong một khoảng cách di truyền ví dụ 250kb). Sau đó nó loại bỏ các SNP gần đó nếu mối tương quan giữa chúng lớn hơn một ngưỡng nhất định (ví dụ r^2 = 0.2, [32]). Như vậy, bước Clumping giúp loại bỏ dữ liệu dư thừa do mất cân bằng liên kết (LD) gây ra. Tập hợp SNP của GWAS tương quan với kiểu hình dưới một ngưỡng P-value nào đó sẽ được lựa chọn để tính toán PRS, phương pháp này còn được gọi là "Đặt ngưỡng" ("Thresholding", T). Vì không thể xác định trước ngưỡng P-value tối ưu, nên nhiều tập SNP với nhiều ngưỡng P-value khác nhau sẽ được lần lượt được đưa vào mô hình huấn luyện. Mô hình này còn có sự đóng góp của các hiệp biến và các thành phần chính của tập đích được tính toán dựa trên phân tầng quần thể. Ngưỡng P-value nào cho ra được độ chính xác dự đoán cao nhất thì tập hợp các SNP tương ứng với ngưỡng đó sẽ được lựa chọn cho tính toán PRS [24, 31, 45].

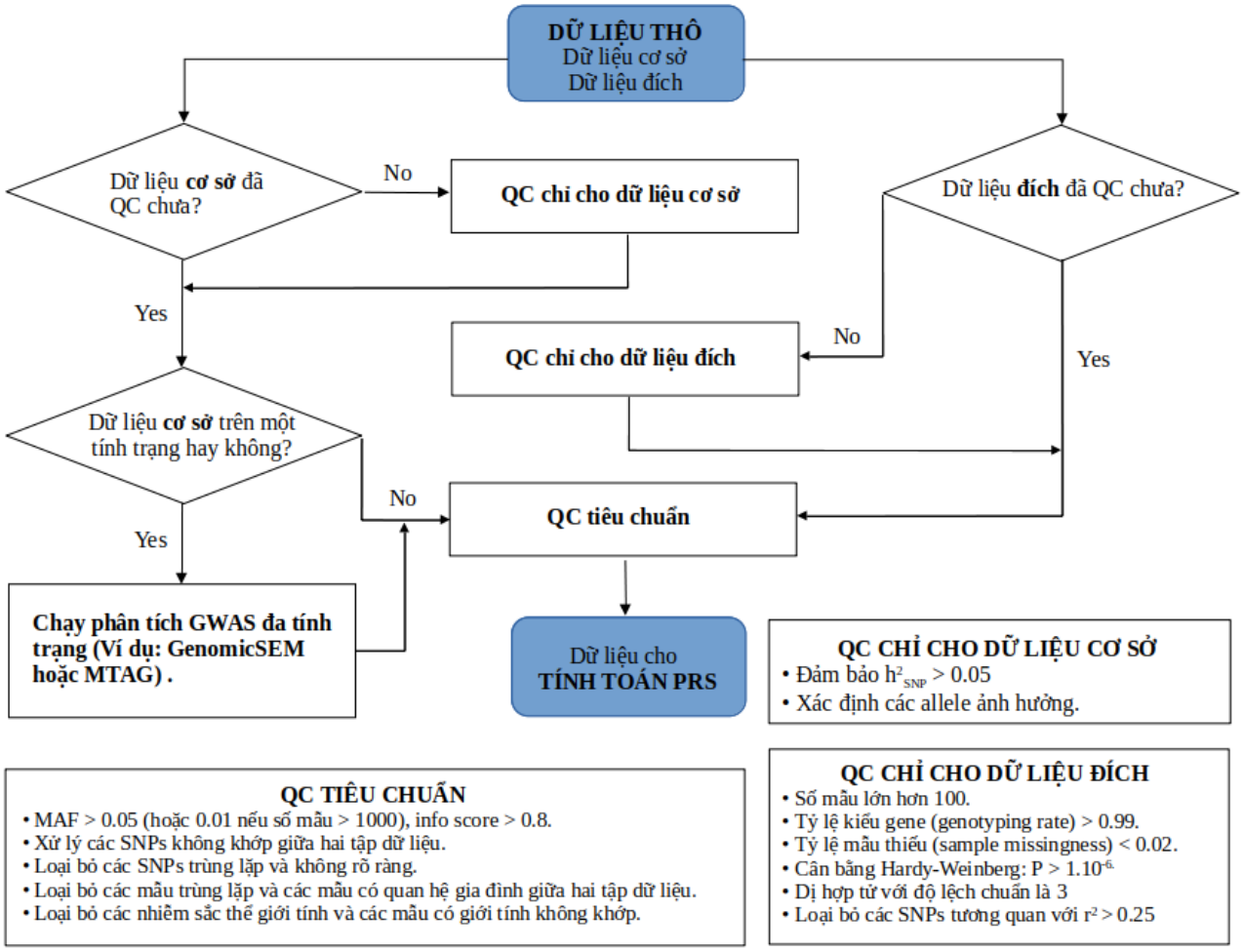
Ưu điểm của phương pháp "Nhóm và Đặt ngưỡng" là tốc độ tính toán nhanh do các tập dữ liệu tương ứng với các ngưỡng khác nhau có dạng ma trận (các hàng là các cá thể, các cột là các SNP) được chuyển về một vec tơ mà mỗi phần tử là điểm nguy cơ đa di truyền của từng cá thể. Tuy nhiên, nhược điểm lớn của phương pháp này là việc lựa chọn các ngưỡng hoàn toàn dựa trên kinh nghiệm, vì vậy khó tìm ra một cách đặt ngưỡng nhất quán cho các nghiên cứu khác nhau. Hơn nữa, phương pháp vẫn có thể bỏ sót các SNP ngoài phạm vi của ngưỡng nhưng thực sự ảnh hưởng đến bệnh.

Thay vì dựa vào kinh nghiệm để lựa chọn các tập SNP, một loạt các phương pháp thu gọn sử dụng học máy đã được phát triển để có thể loại bỏ các SNP không ảnh hưởng đến kiểu hình ngay trong quá trình học. Ngoài ra, các phương pháp mới nhất đi sâu vào việc tìm hiểu các đặc điểm sinh học kết hợp với học máy để giữ lại những SNP thực sự có ảnh hưởng đến bệnh cho dù chúng có thể không độc lập với nhau. Một số phương pháp cũng góp phần cải thiện kiểm soát mất cân bằng liên kết (Xem Phần IV).

IV. NHỮNG CẢI TIẾN SỬ DỤNG HỌC MÁY ĐỂ NÂNG CAO KHẢ NĂNG ÁP DỤNG TRONG LÂM SÀNG CỦA DỰ ĐOÁN NGUY CƠ ĐA DI TRUYỀN

Mặc dù PRS được tạo ra từ các nghiên cứu hệ gen quy mô lớn, nhưng do sự phức tạp của yếu tố đa di truyền kết hợp với các yếu tố môi trường mà việc áp dụng lâm sàng ở quy mô rộng của PRS đối mặt với những thách thức đáng kể. Để nâng cao khả năng sử dụng của PRS trong lâm sàng các nhà khoa học đã đưa ra một số cách tiếp cận khác nhau. Cách tiếp cận thứ nhất, dự đoán chỉ dựa trên các SNP có tương quan cao với bệnh đã biết trong dữ liệu GWAS. Cách tiếp cận thứ hai, các phương pháp chọn biến và học máy được phát triển để thu hẹp tập SNP [46] sử dụng các kỹ thuật điều chỉnh/thu hẹp trong thống kê như LASSO hoặc hồi quy Ridge (Ridge Regression) [47], hoặc sử dụng cách tiếp cận Bayes thông qua việc xác định phân phối [46, 48]. Cách tiếp cận thứ ba, mô hình dự đoán tính đến tỷ lệ của các biến dị nhân quả trong tổng số biến dị di truyền như phương pháp LDpred. Hiệu năng của các mô hình được đánh giá trong Nghiên cứu Dịch tễ học Di truyền về Sức khỏe Người trưởng thành và Lão hóa (the Genetic Epidemiology Research on Adult Health and Aging, GERA) [49] được giới thiệu ở Bảng III.

Hiện nay, các phương pháp chú trọng hơn đến việc xác định các biến dị ảnh hưởng thực sự đến kiểu hình và tìm cách đánh trọng số phù hợp cho các loại biến dị như: cây hồi quy tăng cường gradient và điều chỉnh mất cân bằng liên kết (GrabBLD) [50], dự đoán di truyền đa biến với ngưỡng trơn [51], xác định các điểm đánh dấu di truyền [52]. Mặt khác, điểm nguy cơ đa di truyền thường cung cấp một độ đo tương đối về nguy cơ được đánh giá ở cấp độ một nhóm người chứ không phải ở cấp độ cá nhân [53]



Hình 3. **Sơ đồ kiểm soát chất lượng cho tính toán PRS**. Trước tiên, dữ liệu đích và dữ liệu cơ sở được tiến hành QC riêng, nếu dữ liệu cơ sở được xây dựng trên nhiều tính trạng thì ta cần chạy phân tích GWAS đa tính trạng. Sau đó, cả hai tập dữ liệu được QC tiêu chuẩn để tạo ra một tập dữ liệu phục vụ cho tính toán PRS.

nên việc áp dụng PRS trong lâm sàng như một công cụ dự đoán khả năng mắc bệnh của cá nhân gặp những khó khăn nhất định. Để giải quyết vấn đề này, phương pháp ước lượng giới hạn tin cậy đã cho phép tính toán giá trị xác suất có điều kiện của tình trạng bệnh trên từng cá nhân [53]. Đặc biệt, phương pháp học sâu cũng cho thấy sức mạnh trong việc dự đoán nguy cơ mắc bệnh đa di truyền khi so sánh với một loạt các phương pháp học máy trước đó [54]. Trong khuôn khổ bài báo, chúng tôi sẽ giới thiệu chi tiết hơn một vài phương pháp điển hình gần đây sử dụng học máy được liệt kê trong Bảng IV.

1. Mô hình hồi quy logistic phạt (Penalized Logistic Regression, PLR)

Nhóm nghiên cứu của Florian Privé đã đưa ra mô hình hồi quy logistic phạt (Penalized Logistic Regression, PLR) để cải thiện hiệu quả cho tính toán PRS [55]. Trong mô hình này, hàm điều chỉnh L2 ("Ridge") có tác dụng thu nhỏ các hệ số và hàm điều chỉnh L1 ("LASSO") đưa các một phần các hệ số về giá trị 0 và có thể được sử dụng để chọn biến ngay trong quá trình học. Kết hợp giữa các hàm điều chỉnh L1 và L2 ("Elastic-Net") rất hiệu quả trong trường hợp số lượng SNP lớn hơn rất nhiều số lượng mẫu.

Cụ thể, bài toán được đưa về ước lượng các hệ số β_0, β để cực tiểu hóa hàm tổn thất được điều chỉnh

$$L(\lambda,\alpha) = -\sum_{i=1}^{n}(y_i log(z_i) + (1-y_i)log(1-z_i)) + \lambda((1-\alpha)\frac{1}{2}||\beta||_2^2 + \alpha||\beta||_1)$$

trong đó $z_i = 1/(1+e^{-(\beta_0 + x_i^T\beta)})$, x biểu diễn kiểu gen và các hiệp biến (ví dụ: các thành phần chính, tuổi, giới tính), y là tình trạng bệnh, λ và α là hai siêu tham số điều chỉnh.

Bảng III
SO SÁNH HIỆU NĂNG CỦA CÁC PHƯƠNG PHÁP THUỘC BA CÁCH TIẾP CẬN KHÁC NHAU TRONG TÍNH TOÁN PRS VỚI DỮ LIỆU GERA. BẢNG SO SÁNH ĐƯỢC THAM KHẢO TỪ NGHIÊN CỨU CỦA M. THOMAS VÀ CÁC ĐỒNG NGHIỆP [49].

Các phương pháp tính PRS		**Số biến dị**	**AUC (95% CI)**
Cách tiếp cận 1: Các biến dị GWAS đã biết			
Các biến dị đã biết		140	0.629 (0.613–0.645)
Cách tiếp cận 2: Lựa chọn SNP và Học máy			
Ridge		10000	0.633 (0.617–0.648)
Lasso		10000	0.629 (0.601–0.646)
Elastic Net		10000	0.630 (0.612–0.641)
XGBoost		10000	0.629 (0.614–0.643)
Cách tiếp cận 3: LDpred			
LDpred	ρ = 1	1180765	0.620 (0.603–0.637)
	ρ = 0.3	1180765	0.625 (0.608–0.642)
	ρ = 0.1	1180765	0.628 (0.611–0.645)
	ρ = 0.03	1180765	0.635 (0.619–0.651)
	ρ = 0.01	1180765	0.646 (0.630–0.662)
	ρ = 0.005	1180765	0.649 (0.633–0.664)
	ρ = 0.003	1180765	0.654 (0.639–0.669)
	ρ = 0.001	1180765	0.643 (0.628–0.658)
Với LDpred, ρ là tỉ lệ các biến dị nhân quả trong tổng số biến dị của bệnh ung thư đại trực tràng (Accurate colorectal cancer, CRC).			

Dữ liệu được sử dụng là kiểu gen thực tế của các cá thể người châu Âu từ nghiên cứu dựa trên thuần tập ca bệnh và thuần tập đối chứng cho bệnh Celiac [56]. Để loại bỏ sự phân tầng di truyền gây ra bởi phân tầng quần thể, các tác giả chỉ sử dụng 15,155 cá thể với 4,496 mẫu ca bệnh và 10,659 mẫu đối chứng sau khi lọc; cả hai thuần tập này chứa 281,122 SNP.

Nhóm nghiên cứu đã xây dựng ba kịch bản mô phỏng để xét đến sự ảnh hưởng của các yếu tố đa di truyền cũng như cỡ mẫu đến kết quả dự đoán. PLR cho thấy hiệu quả dự đoán tốt hơn "C+T", giá trị AUC (xem hộp 1) cao hơn, trong hầu hết các kịch bản ngoại trừ một vài trường hợp yếu tố đa di truyền cao và hệ số di truyền thấp. Ngoài ra, nhóm nghiên cứu cũng đã phát triển hai gói công cụ trên R là bigstatsr và bigsnpr. Vấn đề bộ nhớ được giải quyết bằng cách sử dụng định dạng dữ liệu được lưu trữ dưới dạng tệp nhị phân. Các chức năng được cung cấp trong các gói công cụ này đều được song song hóa. Trong khi gói bigstatsr cung cấp các thuật toán tiêu chuẩn trong phân tích thống kê thì gói bigsnpr được xây dựng trên gói bigstatsr cung cấp các thuật toán dành riêng cho GWAS.

2. Cây hồi quy tăng cường gradient và điều chỉnh mất cân bằng liên kết (Gradient Boosted and LD, GraBLD)

Một phương pháp mới sử dụng các kỹ thuật học máy dựa trên cây hồi quy tăng cường gradient và điều chỉnh mất cân bằng liên kết (Gradient Boosted and LD, GraBLD) để tăng hiệu quả dự đoán nguy cơ đa di truyền. Đây là lần đầu tiên mà cây hồi quy tăng cường Gradient được sử dụng để tối ưu hóa trọng số của SNP kết hợp với việc điều chỉnh các vùng gây ra bởi sự mất cân bằng liên kết [50]. Một số kế quả khả quan của GraBLD áp dụng với các bộ dữ liệu tương ứng với các tính trạng khác nhau được liệt kê dưới đây:

- Dữ liệu thống kê tóm tắt từ Điều tra di truyền các tính trạng nhân trắc học (Genetic Investigation of Anthropometric Traits, GIANT) cho tính trạng chiều cao và tính trạng trọng lượng cơ thể trên bình phương chiều cao (Body Mass Index, BMI) trong nghiên cứu của UK Biobank [57] (130 nghìn cá thể; 1.98 triệu SNP): GrabBLD giải thích được tương ứng 46.9% và 32.7% của toàn bộ phương sai đa di truyền.
- Dữ liệu thống kê tóm tắt từ phân tích tổng hợp và sao chép di truyền bệnh tiểu đường (Diabetes Genetics Replication And Meta-analysis, DIAGRAM) trong nghiên cứu UK Biobank: diện tích dưới đường cong đặc tính (AUC) là 0.602.

GraBLD vượt trội hơn các phương pháp tính điểm truyền thống khi áp dụng để dự đoán chiều cao ($p < 2.2 * 10^{-16}$) và BMI ($p < 1.57 * 10^{-4}$), và tương đương với LDpred đối với bệnh tiểu đường. Về mặt phương pháp, cây hồi quy tăng cường gradient là phương pháp mạnh mẽ và linh hoạt kết hợp các phân loại yếu để tạo ra một phương pháp học mạnh cho dự đoán đầu ra là biến liên tục. GraBLD có khả năng cải thiện các trọng số SNP trong điểm nguy cơ đa di truyền mà không yêu cầu kiểu gen ở mức độ cá thể vì nó có thể mô hình hóa các mối quan hệ phi tuyến mà không cần lựa chọn đặc trưng.

3. Dự đoán di truyền đa biến với ngưỡng trơn (Smooth-Threshold Multivariate Genetic Prediction, STMGP)

Nghiên cứu của Yuta Takahashi và các đồng nghiệp vào năm 2020 đã đưa ra một thuật toán dự đoán mới, dự đoán di truyền đa biến với ngưỡng trơn (Smooth-Threshold Multivariate Genetic Prediction, STMGP), thuật toán đã cải thiện độ chính xác trong việc dự đoán các kiểu hình bệnh thần kinh đa di truyền bằng cách giảm overfitting (Xem định nghĩa trong hộp 1) thông qua việc lựa chọn các biến dị và xây dựng mô hình hồi quy phạt [51]. Mô hình dự đoán sử dụng tập huấn luyện gồm 3685 người ở quận Miyagi, tập kiểm thử độc lập gồm 3048 người ở quận Iwate tại Nhật Bản. Trong đó, kiểu hình là các triệu chứng trầm cảm và các kiểu hình được mô phỏng với độ phức tạp khác nhau và sự phân bố mức độ ảnh hưởng khác nhau của các alen nguy cơ (xem định nghiã trong hộp 1).

Nghiên cứu đã chỉ ra nguyên nhân khiến các nghiên cứu di truyền trước đây bị hạn chế trong việc dự đoán các kiểu hình của bệnh tâm thần bằng cách xem xét hai loại biến dị chủ yếu (Xem Hình 4):

Bảng IV
CÁC NGHIÊN CỨU NỔI BẬT GẦN ĐÂY SỬ DỤNG HỌC MÁY ĐỂ CẢI THIỆN ĐỘ CHÍNH XÁC CỦA DỰ ĐOÁN BỆNH ĐA DI TRUYỀN.

STT	Tên nghiên cứu	Tác giả	Năm	Phương pháp	Nội dung chính
1	Machine-learning heuristic to improve gene score prediction of polygenic traits	G. Paré	2017	GraBLD	Sử dụng cây hồi quy tăng cường gradient và điều chỉnh mất cân bằng liên kết để tăng hiệu quả dự đoán nguy cơ mắc bệnh đa di truyền.
2	Efficient Implementation of Penalized Regression for Genetic Risk Prediction	F. Privé	2019	PLR	Sử dụng mô hình hồi quy logistic phạt với dữ liệu kiểu gen-kiểu hình và các hiệp biến để chọn các đặc trưng ngay trong quá trình học.
3	Machine learning for effectively avoiding overfitting is a crucial strategy for the genetic prediction of polygenic psychiatric phenotypes	Y. Takahashi	2020	STMGP	Phương pháp giúp giữ lại các biến dị nhạy thực sự có ảnh hưởng đến bệnh nhưng có tương quan với nhau, và loại bỏ các biến dị rỗng bằng việc đánh trọng số các tập biến dị và xây dựng hồi quy ridge.
4	Translating polygenic risk scores for clinical use by estimating the confidence bounds of risk prediction	J. Sun	2021	MCCP	Dựa vào độ đo NCM được áp dụng trên tập hiệu chỉnh để tìm xác suất mà một cá thể thuộc về một lớp, từ đó xác định được giới hạn tin cậy của một dự đoán.
5	Deep neural network improves the estimation of polygenic risk scores for breast cancer	A. Badré	2021	DNN	Mô hình được chứng minh là hoạt động tốt hơn các kỹ thuật học máy và các thuật toán thống kê truyền thống như BLUP, BayesA và Ldpred nhờ tạo ra được hai phân phối chuẩn của hai lớp có trung bình khác biệt và tính đến các ảnh hưởng phi tuyến.
6	Improving the Utility of Polygenic Risk Scores as a Biomarker for Alzheimer's Disease	D. Vlachakis	2021	Biomarker	Xây dựng một quy trình lọc và đánh trọng số dữ liệu SNP, tạo tiền đề cho việc áp dụng các kỹ thuật học máy trong việc nâng cao độ chính xác của dự đoán AD, cũng như các bệnh đa di truyền khác.

- Biến dị nhạy thực sự: là các biến dị ảnh hưởng đến bệnh. Trong đó, có các biến dị nhạy thực sự độc lập với nhau (màu cam) và các biến dị nhạy thực sự tương quan với nhau (màu vàng). Các biến dị này làm tăng độ chính xác của dự đoán.
- Biến dị rỗng (màu xanh): là các biến dị không ảnh hưởng đến bệnh. Các biến dị rỗng làm giảm độ chính xác của dự đoán.

Do khả năng giới hạn trong việc phân biệt các biến dị nhạy thực sự với các biến dị rỗng, nên mô hình có thể khớp một số lượng lớn nhiễu là các biến dị rỗng, kết quả dẫn đến overfitting (Xem định nghĩa trong hộp 1). Đây chính là vấn đề của các nghiên cứu về PRS trước đó khi nó ước lượng quá mức một số lượng lớn các biến dị rỗng kết hợp với các biến dị được nhóm trong quá trình Clumping (Xem Phần III.2), cũng như loại trừ các biến dị thực sự nhạy tương quan với nhau nhưng có liên quan đến bệnh.

STMGP xây dựng mô hình dự đoán dựa trên tập các biến dị được lựa chọn bởi các ngưỡng P-value của GWAS tương tự như PRS. Tuy nhiên, STMGP khắc phục được các vấn đề của PRS ở hai khía cạnh. Thứ nhất, STMGP có thể tránh overfitting bởi việc đánh trọng số các biến dị, trong đó các biến dị nhạy thực sự có trọng số cao hơn các biến dị rỗng. Thứ hai, STMGP sử dụng cả các biến dị nhạy thực sự tương quan với nhau mà có thể đóng góp vào độ chính xác dự đoán bởi việc xây dựng hồi quy Ridge tổng quát.

4. Ước lượng giới hạn tin cậy của dự đoán nguy cơ đa di truyền

Để dự đoán tình trạng bệnh cho từng cá nhân Jiangming Sun và đồng nghiệp đã giới thiệu một kỹ thuật học máy, MCCP (Mondrian Cross-Conformal Prediction), để tìm ra giá trị xác suất có điều kiện về tình trạng bệnh cho từng cá nhân thông qua ước lượng giới hạn tin cậy của dự đoán PRS tới nguy cơ mắc bệnh [53]. MCCP xây dựng độ đo NCM (Nonconformity Measures) nhờ các giá trị được dự đoán trên tập hiệu chỉnh bằng mô hình được huấn luyện trên tập huấn luyện (độc lập và cùng phân phối với tập hiệu chỉnh); tập huấn luyện và tập hiệu chỉnh đều chứa thông tin kiểu hình.

$$NCM_y = -y * d(x_i)$$

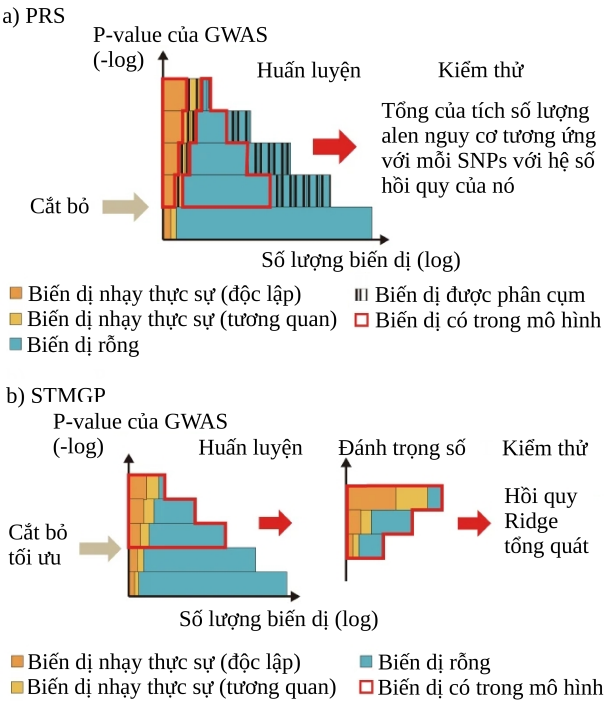
trong đó y là các lớp khác không, chẳng hạn (1, -1), và $d(x_i)$ là giá trị quyết định nhận được từ hàm quyết định của mô hình được học từ dữ liệu huấn luyện. Xác suất mà cá thể i thuộc về lớp y được tính như sau:

$$p_y^i = \frac{|\{j = 1, ..., N_{hc_y} : y_i = y, NCM_j \geq NCM_i\}|}{\{N_{hc_y} + 1 : y_i = y\}}$$

trong đó N_{hc_y} là cỡ mẫu của lớp y trong tập hiệu chỉnh. Với một sai số kỳ vọng $\alpha \in [0, 1]$ kết quả của vùng dự đoán được xác định dưới đây:

$$\Gamma^\alpha = \{y \in Y : p_y > \alpha\}$$

trong đó Y là tập hợp các lớp, p_y là giá trị xác suất khi một cá thể thuộc về lớp y, vùng dự đoán Γ^α là một tập



Hình 4. **So sánh giữa PRS và STMGP**. a) PRS bị overfitting do mô hình khớp với một số lượng lớn nhiễu là các biến dị rỗng (màu xanh) và các biến dị được nhóm bởi Clumping. Trong khi nó chỉ giữ lại các biến dị nhạy thực sự độc lập (màu cam) và loại bỏ các biến dị nhạy thực sự có tương quan với nhau (màu vàng) nhưng lại đóng góp vào độ chính xác của mô hình. b) STMGP không những tránh được overfitting do việc đánh trọng số các biến dị, mà còn giữ lại được các biến dị nhạy thực sự có tương quan với nhau nhưng ảnh hưởng đến bệnh nhờ mô hình hồi quy ridge tổng quát. Hình ảnh được tham khảo từ nghiên cứu của Yuta Takahashi và các đồng nghiệp [51].

hợp có thể rỗng hoặc chứa một hoặc chứa hai lớp. Độ tin cậy của MCCP được định nghĩa như sau:

$$sup\{1-\alpha : \Gamma^{\alpha} \leq 1\}$$

là giá trị lớn nhất của $1-\alpha$ với Γ^{α} là một lớp duy nhất, chẳng hạn lớp bệnh hoặc lớp đối chứng. Như vậy, với một cá thể có trạng thái được dự đoán, MCCP có thể ước lượng giới hạn tin cậy của dự đoán đó.

Nhóm nghiên cứu đã áp dụng MCCP đối với bệnh động mạch vành (CAD), đái tháo đường týp 2 (T2D), bệnh viêm ruột (IBD) và ung thư vú (*BRCA*) sử dụng tài nguyên của UK Biobank [9, 57] và hai bộ dữ liệu bổ sung dựa trên quần thể, mẫu bệnh tâm thần phân liệt (SCZ) trong nghiên cứu tâm thần tích hợp (iPSYCH) [58] và mẫu T2D trong nghiên cứu chế độ ăn kiêng Malmo và ung thư (MDC) [59]. Kết quả cho thấy ở cấp độ cá nhân, MCCP báo cáo xác suất dự đoán được hiệu chỉnh tốt, ước tính có hệ thống giới hạn tin cậy của của PRS trong dự đoán nguy cơ mắc bệnh phức tạp ở người. Ở cấp độ nhóm, MCCP vượt trội hơn các phương pháp tiêu chuẩn trong việc phân tầng chính xác các cá nhân thành các nhóm nguy cơ.

5. Mạng nơ-ron sâu (Deep Neural Network, DNN)

Vào năm 2021, Adrien Badré và đồng nghiệp đã so sánh một loạt các mô hình tính toán để ước tính PRS cho bệnh ung thư vú. Một mạng nơ-ron sâu (Deep Neural Network, DNN) được chứng minh là hoạt động tốt hơn các kỹ thuật học máy và các thuật toán thống kê truyền thống như BLUP, BayesA và LDpred [54]. Về hiệu quả của mô hình, diện tích dưới đường cong đặc tính (AUC) là 67,4% đối với DNN, 64,2% đối với BLUP, 64,5% đối với BayesA, và 62,4% đối với LDpred. Ưu điểm lớn nhất của DNN là có thể tách quần thể ca bệnh thành hai quần thể: một quần thể con có nguy cơ di truyền cao với PRS trung bình cao hơn đáng kể so với quần thể đối chứng, một quần thể con có nguy cơ bình thường với PRS trung bình tương tự như quần thể đối chứng. Nguyên nhân là vì PRS do DNN tạo ra tuân theo hai phân phối chuẩn với các trung bình khác nhau rõ rệt. Ngược lại, BLUP, BayesA và LPpred đều tạo ra PRS tuân theo chỉ một phân phối chuẩn trong quần thể ca bệnh. Điều này cho phép DNN đạt được recall 18,8% ở precision 90% (Xem định nghĩa recall và precision ở hộp 1) trong thuần tập kiểm thử. Mặt khác, mô hình DNN còn xác định được các biến dị đặc trưng mà chỉ được gán giá trị P-value không đáng kể bởi GWAS, các biến dị này có thể tương quan với kiểu hình thông qua các mối quan hệ phi tuyến. Đây cũng là ưu điểm lớn của các mạng học sâu so với các phương pháp học máy truyền thống.

6. Xác định các điểm đánh dấu di truyền

Để cải thiện độ đo PRS, các nhà khoa học đã đề xuất một quy trình mới áp dụng cho bệnh Alzheimer (AD) nhằm nâng cao độ chính xác dự đoán từ các yếu tố đa di truyền [52]. Nhóm tác giả lạc quan rằng các phân tích dự kiến trong tương lai của họ sẽ chứng minh cách tiếp cận này có thể nâng cao đáng kể hiệu suất của PRS và khả năng ứng dụng của nó trong lâm sàng. Các bước của quy trình được tóm tắt như sau:

- Bước 1: Xem xét mạng quan hệ giữa các SNP trong các khu vực gần gen AD. Có thể có các SNP không ảnh hưởng đến kiểu hình đi kèm với các SNP đặc chưng vì chúng được liên kết hoặc di truyền cùng nhau.
- Bước 2: Đánh trọng số các SNP cao hơn trong các gen mã hóa vì những thay đổi mà chúng mang lại cho chuỗi polypeptit có thể có tác động sửa đổi sau dịch mã.
- Bước 3: Đánh trọng số cao hơn cho các SNP tương quan với các dạng nghiêm trọng hơn của kiểu hình bắt nguồn từ các nghiên cứu GWAS khác nhau.

Các nhà khoa học đã phát triển một tập dữ liệu tích hợp với tất cả các gen liên quan đến bệnh Alzheimer. Các SNP từ các nghiên cứu được ánh xạ vào 1241 gen liên quan đến AD mà được lựa chọn theo qui trình. Kết quả cuối cùng cho ra ba nhóm SNP. Một nhóm bao gồm các SNP được xác định trong cả tập dữ liệu bệnh Alzheimer và bệnh Alzheimer khởi phát muộn mà không được thưởng hay phạt. Một nhóm bao gồm các SNP chỉ có trong các nghiên cứu GWAS về AD và liên tiếp được thưởng (vì chúng thuộc nhóm biểu hiện lâm sàng nặng). Một nhóm bao gồm các các SNP chỉ có trong dữ liệu AD khởi phát muộn và bị phạt vì chúng được liên kết với một biểu hiện nhẹ hơn của AD. Phương pháp này có thể mở đường cho các ứng dụng mới của học máy hệ gen vào dữ liệu GWAS trong nỗ lực xác định các điểm đánh dấu di truyền cho các bệnh phức tạp.

V. CÁC XU HƯỚNG CẢI TIẾN HIỆU NĂNG CHO MÔ HÌNH DỰ ĐOÁN NGUY CƠ ĐA DI TRUYỀN TRONG TƯƠNG LAI

1. Mở rộng nghiên cứu tương quan toàn hệ gen

Với sự phát triển mạnh mẽ của các nghiên cứu tương quan toàn hệ gen, GWAS sẽ không chỉ tập trung vào nhóm người châu Âu mà đang mở rộng ra các nhóm người khác trên toàn thế giới [60–66]. Điều này làm cho dữ liệu GWAS trở nên đáng tin cậy hơn khi sử dụng để dự đoán nguy cơ đa di truyền. Hơn nữa, các cơ sở dữ liệu như HapMap [67], dự án 1000 hệ gen [68] vẫn đang tiếp tục được hoàn thiện giúp cho quá trình suy diễn thống kê bổ sung các SNP bị thiếu trong các bộ dữ liệu từ các nghiên cứu riêng lẻ trở nên chính xác hơn [69, 70]. Các phân tích tổng hợp (meta analysis) kết hợp các nghiên cứu riêng lẻ với nhau sẽ tạo nên những bộ dữ liệu lớn bao gồm nhiều tính trạng khác nhau [71–74]. Đặc biệt, xu hướng kết hợp dữ liệu GWAS với dữ liệu giải trình tự thế hệ mới đang trở nên phổ biến khi mà giá thành giải trình tự ngày càng rẻ [75, 76]. Điều này góp phần tạo ra một số lượng SNP lớn cần thiết mà trước đây chưa bao giờ có.

Như vậy, nhờ sự phát triển mạnh mẽ các nghiên cứu mà các bộ dữ liệu GWAS ngày càng đầy đủ và đáng tin cậy. Đây là điều kiện cần để thúc đẩy các nghiên cứu dự đoán nguy cơ đa di truyền cho nhiều tính trạng mới với độ chính xác cao hơn trước, từ đó giúp mở rộng phạm vi áp dụng của PRS trong lâm sàng.

2. Phát triển mô hình dự đoán nguy cơ mắc bệnh

Để hướng tới y học chính xác, các mô hình dự đoán nguy cơ đa di truyền đang dần được phát triển để có thể xác định nguy cơ mắc bệnh của người trong một khung thời gian cụ thể. Đại diện cho hướng phát triển này là mô hình tính toán nguy cơ tuyệt đối được giới thiệu bởi Nilanjan Chatterjee và các cộng sự vào năm 2016 [77]. Từ đó đến nay đã có một số nghiên cứu về nguy cơ tuyệt đối cho các bệnh khác nhau như: động mạch vành [78], ung thư vú [79], ung thư phổi [80], ung thư đại trực tràng [81], ung thư tuyến tiền liệt [82]. Để cải thiện hiệu năng dự đoán, ta có thể bổ sung các yếu tố được tổng hợp từ các nghiên cứu dịch tễ chất lượng cao với cỡ mẫu lớn bao gồm: biến dị di truyền, yếu tố nguy cơ từ môi trường, các dấu hiệu sinh học về phơi nhiễm, trong đó có tính đến sự tương tác giữa các yếu tố nguy cơ. Đặc biệt để dự báo nguy cơ phát triển của bệnh trong một khoảng thời gian cụ thể ta cần thêm phân phối của các yếu tố nguy cơ như độ tuổi khởi phát bệnh, tỷ lệ tử vong trong quần thể nghiên cứu. Khi đó mô hình có khả năng tính toán xác suất của một cá thể không có triệu chứng sẽ phát triển bệnh trong một khoảng thời gian nào đó.

Việc phát triển mô hình dự đoán nguy cơ đa di truyền nhằm xác định nguy cơ tuyệt đối có tiềm năng lớn trong tương lai, nó có thể được áp dụng rộng rãi trong lâm sàng khi khả năng tiếp cận các dữ liệu lâm sàng và dữ liệu dịch tễ trở nên dễ dàng hơn.

VI. KẾT LUẬN

Trong bài báo này chúng tôi đã giới thiệu một cách tổng quan về dự đoán nguy cơ đa di truyền từ quá trình phát triển của GWAS cho đến các nghiên cứu cải tiến phương pháp tính toán PRS liên quan đến học máy.

Chúng tôi đã phân tích và sắp xếp lại các bước kiểm soát chất lượng để đưa ra một quy trình tiền xử lý dữ liệu rõ ràng và đầy đủ. Đây là quy trình rất quan trọng vì nó ảnh hưởng trực tiếp đến hiệu suất của mô hình khi mà dữ liệu phục vụ cho quá trình huấn luyện mô hình thường đến từ nhiều nghiên cứu khác nhau. Mặt khác, chúng tôi cũng thấy được những thách thức mà các nhà khoa học gặp phải khi tìm cách áp dụng PRS vào lâm sàng. Vì vậy, những nhược điểm của PRS và các phương pháp ngày càng tốt hơn giúp giải quyết những nhược điểm đó cũng được liệt kê một cách khá đầy đủ. Theo đó, một loạt các nghiên cứu cải tiến điển hình liên quan đến học máy trong thời gian gần đây được lựa chọn và trình bày tóm tắt một cách có hệ thống.

Với đa dạng các nghiên cứu theo nhiều hướng tiếp cận khác nhau của học máy đã góp phần nâng cao độ chính xác của các dự đoán dựa vào PRS cho các bệnh phổ biến. Các tác giả không những tập trung vào loại bỏ các SNP là nhiễu do mất cân bằng liên kết mà còn xác định được chính xác hơn ảnh hưởng của các SNP với kiểu hình. Các phương pháp xác định có thể tìm được các ảnh hưởng tuyến tính như tương quan giữa SNP với kiểu hình, cũng có thể tìm được các ảnh hưởng phi tuyến. Độ chính xác của PRS

được cải thiện góp phần nâng cao độ chính xác của tính toán nguy cơ tuyệt đối giúp phân tầng nguy cơ di truyền trong lâm sàng.

Chúng tôi hy vọng cái nhìn tổng quan về PRS, nhất là trong thời gian gần đây, sẽ giúp các nhà nghiên cứu dễ dàng hơn trong việc tiếp cận, cũng như tìm ra những hướng nghiên cứu cải tiến mới có giá trị nhằm mở rộng phạm vi ứng dụng của PRS trong lâm sàng.

LỜI CẢM ƠN

Toàn bộ nghiên cứu được tài trợ bởi quỹ Nghiên cứu và Ứng dụng LB.Sci của Công ty TNHH LOBI Việt Nam.

TÀI LIỆU THAM KHẢO

[1] Kuchenbaecker, K.B. et al. (2017) "Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers", *JAMA*, 317(23), pp. 2402–2416. doi:10.1001/jama.2017.7112.

[2] Wexler, N.S. et al. (1987) "Homozygotes for Huntington's disease", *Nature*, 326(6109), pp. 194–197. doi:10.1038/326194a0.

[3] Gusella, J.F. (1989) "Location cloning strategy for characterizing genetic defects in Huntington's disease and Alzheimer's disease", *FASEB journal: official publication of the Federation of American Societies for Experimental Biology*, 3(9), pp. 2036–2041. doi:10.1096/fasebj.3.9.2568302.

[4] Ford, D. et al. (1998) "Genetic Heterogeneity and Penetrance Analysis of the BRCA1 and BRCA2 Genes in Breast Cancer Families", *The American Journal of Human Genetics*, 62(3), pp. 676–689. doi:10.1086/301749.

[5] Klein, R.J. et al. (2005) "Complement factor H polymorphism in age-related macular degeneration", Science (New York, N.Y.), 308(5720), pp. 385–389. doi:10.1126/science.1109557.

[6] Burton, P.R. et al. (2007) "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls", *Nature*, 447(7145), pp. 661–678. doi:10.1038/nature05911.

[7] Loos, R.J.F. (2020) "15 years of genome-wide association studies and no signs of slowing down", *Nature Communications*, 11(1), p. 5900. doi:10.1038/s41467-020-19653-5.

[8] Visscher, P.M. et al. (2017) "10 Years of GWAS Discovery: Biology, Function, and Translation", *American Journal of Human Genetics*, 101(1), pp. 5–22. doi:10.1016/j.ajhg.2017.06.005.

[9] Bycroft, C. et al. (2018) "The UK Biobank resource with deep phenotyping and genomic data", *Nature*, 562(7726), pp. 203–209. doi:10.1038/s41586-018-0579-z.

[10] Pepe, M.S. et al. (2004) "Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker", *American Journal of Epidemiology*, 159(9), pp. 882–890. doi:10.1093/aje/kwh101.

[11] Jakobsdottir, J. et al. (2009) "Interpretation of Genetic Association Studies: Markers with Replicated Highly Significant Odds Ratios May Be Poor Classifiers", *PLOS Genetics*, 5(2), p. e1000337. doi:10.1371/journal.pgen.1000337.

[12] Wray, N.R., Goddard, M.E. and Visscher, P.M. (2007) "Prediction of individual genetic risk to disease from genome-wide association studies", *Genome Research*, 17(10), pp. 1520–1528. doi:10.1101/gr.6665407.

[13] Manolio, T.A. et al. (2009) "Finding the missing heritability of complex diseases", *Nature*, 461(7265), pp. 747–753. doi:10.1038/nature08494

[14] Cecile, A., Janssens, J.W. and Joyner, M.J. (2019) "Polygenic Risk Scores That Predict Common Diseases Using Millions of Single Nucleotide Polymorphisms: Is More, Better?", *Clinical Chemistry*, 65(5), pp. 609–611. doi:10.1373/clinchem.2018.296103.

[15] Sha, Z., Hu, T. and Chen, Y. (2021) "Feature Selection for Polygenic Risk Scores using Genetic Algorithm and Network Science", in *2021 IEEE Congress on Evolutionary Computation (CEC)*, pp. 802–808. doi:10.1109/CEC45853.2021.9504993.

[16] Klinger, J.E. et al. (2021) *Interaction-based feature selection algorithm outperforms polygenic risk score in predicting Parkinson's Disease status*. medRxiv, p. 2021.07.20.21260848. doi:10.1101/2021.07.20.21260848.

[17] Kulm, S., Mezey, J. and Elemento, O. (2022) Benchmarking Polygenic Risk Score Model Assumptions: towards more accurate risk assessment. *bioRxiv*, p. 2022.02.18.480983. doi:10.1101/2022.02.18.480983.

[18] Privé, F. et al. (2019) "Making the Most of Clumping and Thresholding for Polygenic Scores", *The American Journal of Human Genetics*, 105(6), pp. 1213–1221. doi:10.1016/j.ajhg.2019.11.001.

[19] Hahn, G. et al. (2021) "A fast and efficient smoothing approach to Lasso regression and an application in statistical genetics: polygenic risk scores for chronic obstructive pulmonary disease (COPD)", *Statistics and Computing*, 31(3), p. 35. doi:10.1007/s11222-021-10010-0.

[20] Pattee, J. and Pan, W. (2020) "Penalized regression and model selection methods for polygenic scores on summary statistics", *PLOS Computational Biology*, 16(10), p. e1008271. doi:10.1371/journal.pcbi.1008271.

[21] Dickson, S.P. et al. (2021) "GenoRisk: A polygenic risk score for Alzheimer's disease", *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 7(1), p. e12211.

[22] Peng, J. et al. (2021) A Deep Learning-based Genome-wide Polygenic Risk Score for Common Diseases Identifies Individuals with Risk. *medRxiv*, p. 2021.11.17.21265352. doi:10.1101/2021.11.17.21265352.

[23] Zhao, B. and Zou, F. (2021) "On polygenic risk scores for complex traits prediction", *Biometrics* [Preprint]. doi:10.1111/biom.13466.

[24] Euesden, J., Lewis, C.M. and O'Reilly, P.F. (2015a) "PRSice: Polygenic Risk Score software", *Bioinformatics (Oxford, England)*, 31(9), pp. 1466–1468. doi:10.1093/bioinformatics/btu848.

[25] Uffelmann, E. et al. (2021) "Genome-wide association studies", *Nature Reviews Methods Primers*, 1(1), pp. 1–21. doi:10.1038/s43586-021-00056-9.

[26] Purcell, S. et al. (2007) "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses", *American Journal of Human Genetics*, 81(3), pp. 559–575.

[27] Marees, A.T. et al. (2018) "A tutorial on conducting genome-wide association studies: Quality control and statistical analysis", *International Journal of Methods in Psychiatric Research*, 27(2), p. e1608. doi:10.1002/mpr.1608.

[28] Buniello, A. et al. (2019) "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019", *Nucleic Acids Research*, 47(D1), pp. D1005–D1012. doi:10.1093/nar/gky1120.

[29] Tryka, K.A. et al. (2014) "NCBI's Database of Genotypes and Phenotypes: dbGaP", *Nucleic Acids Research*, 42(D1), pp. D975–D979. doi:10.1093/nar/gkt1211.

[30] Sirugo, G., Williams, S.M. and Tishkoff, S.A. (2019) "The Missing Diversity in Human Genetic Studies", *Cell*, 177(1), pp. 26–31. doi:10.1016/j.cell.2019.02.048.

[31] Purcell, S.M. et al. (2009) "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder", *Nature*, 460(7256), pp. 748–752. doi:10.1038/nature08185.

[32] Wray, N.R. et al. (2014) "Research review: Polygenic methods and their application to psychiatric traits", *Journal of Child Psychology and Psychiatry*, and Allied Disciplines, 55(10), pp. 1068–1087. doi:10.1111/jcpp.12295.

[33] Khera, A.V. et al. (2018) "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations", *Nature Genetics*, 50(9), pp. 1219–1224. doi:10.1038/s41588-018-0183-z.

[34] Mavaddat, N. et al. (2019) "Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes", *American Journal of Human Genetics*, 104(1), pp. 21–34. doi:10.1016/j.ajhg.2018.11.002.

[35] Hormozdiari, F. et al. (2015) "Identification of causal genes for complex traits", *Bioinformatics*, 31(12), pp. i206–i213. doi:10.1093/bioinformatics/btv240.

[36] Visscher, P.M., Hill, W.G. and Wray, N.R. (2008) "Heritability in the genomics era — concepts and misconceptions", *Nature Reviews Genetics*, 9(4), pp. 255–266. doi:10.1038/nrg2322.

[37] Lusted, L.B. (1971) "Signal detectability and medical decision-making", *Science (New York, N.Y.), 171(3977)*, pp. 1217–1219. doi:10.1126/science.171.3977.1217.

[38] Fawcett, T. (2006) "An introduction to ROC analysis", *Pattern Recognition Letters*, 27(8), pp. 861–874. doi:10.1016/j.patrec.2005.10.010.

[39] Hanley, J.A. and McNeil, B.J. (1982) "The meaning and use of the area under a receiver operating characteristic (ROC) curve", *Radiology*, 143(1), pp. 29–36. doi:10.1148/radiology.143.1.7063747.

[40] Wray, N.R. et al. (2021) "From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer", *JAMA Psychiatry*, 78(1), pp. 101–109. doi:10.1001/jamapsychiatry.2020.3049.

[41] Anderson, C.A. et al. (2010) "Data quality control in genetic case-control association studies", *Nature Protocols*, 5(9), pp. 1564–1573. doi:10.1038/nprot.2010.116.

[42] Coleman, J.R.I. et al. (2016) "Quality control, imputation and analysis of genome-wide genotyping data from the Illumina HumanCoreExome microarray", *Briefings in Functional Genomics*, 15(4), pp. 298–304. doi:10.1093/bfgp/elv037.

[43] Choi, S.W., Mak, T.S.-H. and O'Reilly, P.F. (2020) "Tutorial: a guide to performing polygenic risk score analyses", *Nature Protocols*, 15(9), pp. 2759–2772. doi:10.1038/s41596-020-0353-1.

[44] Han, B. and Eskin, E. (2011) "Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies", *American Journal of Human Genetics*, 88(5), pp. 586–598. doi:10.1016/j.ajhg.2011.04.014.

[45] Allan, B.L. (1987) "Calculating medication error rates", *American Journal of Hospital Pharmacy*, 44(5), pp. 1044, 1046.

[46] Ge, T. et al. (2019) "Polygenic prediction via Bayesian regression and continuous shrinkage priors", *Nature Communications*, 10(1), p. 1776. doi:10.1038/s41467-019-09718-5.

[47] Mak, T.S.H. et al. (2017) "Polygenic scores via penalized regression on summary statistics", *Genetic Epidemiology*, 41(6), pp. 469–480. doi:10.1002/gepi.22050.

[48] Newcombe, P.J. et al. (2019) "A flexible and parallelizable approach to genome-wide polygenic risk scores", *Genetic Epidemiology*, 43(7), pp. 730–741. doi:10.1002/gepi.22245.

[49] Thomas, M. et al. (2020) "Genome-wide Modeling of Polygenic Risk Score in Colorectal Cancer Risk", *The American Journal of Human Genetics*, 107(3), pp. 432–444. doi:10.1016/j.ajhg.2020.07.006.

[50] Paré, G., Mao, S. and Deng, W.Q. (2017) "A machine-learning heuristic to improve gene score prediction of polygenic traits", *Scientific Reports*, 7(1), p. 12665. doi:10.1038/s41598-017-13056-1.

[51] Takahashi, Y. et al. (2020) "Machine learning for effectively avoiding overfitting is a crucial strategy for the genetic prediction of polygenic psychiatric phenotypes", *Translational Psychiatry*, 10(1), pp. 1–11. doi:10.1038/s41398-020-00957-5.

[52] Vlachakis, D. et al. (2021) "Improving the Utility of Polygenic Risk Scores as a Biomarker for Alzheimer's Disease", *Cells*, 10(7), p. 1627. doi:10.3390/cells10071627.

[53] Sun, J. et al. (2021) "Translating polygenic risk scores for clinical use by estimating the confidence bounds of risk prediction", *Nature Communications*, 12(1), p. 5276. doi:10.1038/s41467-021-25014-7.

[54] Badré, A. et al. (2021) "Deep neural network improves the estimation of polygenic risk scores for breast cancer", *Journal of Human Genetics*, 66(4), pp. 359–369. doi:10.1038/s10038-020-00832-7.

[55] Privé, F., Aschard, H. and Blum, M.G.B. (2019) "Efficient Implementation of Penalized Regression for Genetic Risk Prediction", *Genetics*, 212(1), pp. 65–74. doi:10.1534/genetics.119.302019.

[56] Dubois, P.C.A. et al. (2010) "Multiple common variants for celiac disease influencing immune gene expression", *Nature Genetics*, 42(4), pp. 295–302. doi:10.1038/ng.543.

[57] Sudlow, C. et al. (2015) "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age", *PLOS Medicine*, 12(3), p. e1001779. doi:10.1371/journal.pmed.1001779.

[58] Pedersen, C.B. et al. (2018) "The iPSYCH2012 case–cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders", *Molecular Psychiatry*, 23(1), pp. 6–14. doi:10.1038/mp.2017.196.

[59] Berglund, G. et al. (1993) "The Malmo Diet and Cancer Study. Design and feasibility", *Journal of Internal Medicine*, 233(1), pp. 45–51. doi:10.1111/j.1365-2796.1993.tb00647.x.

[60] Stevenson, A. et al. (2019) "Neuropsychiatric Genetics of African Populations-Psychosis (NeuroGAP-Psychosis): a case-control study protocol and GWAS in Ethiopia, Kenya, South Africa and Uganda", *BMJ Open*, 9(2), p. e025469. doi:10.1136/bmjopen-2018-025469.

[61] Wang, Y.-F. et al. (2021) "Multi-ancestral GWAS identifies shared and Asian-specific loci for SLE and links type III interferon signaling and lysosomal function to the disease", *Arthritis & Rheumatology (Hoboken, N.J.)* [Preprint]. doi:10.1002/art.42021.

[62] Swart, Y. et al. (2022) GWAS in the southern African context. *bioRxiv*. doi:10.1101/2022.02.16.480704.

[63] Shen, H. et al. (2020) "Polygenic prediction and GWAS of depression, PTSD, and suicidal ideation/self-harm in a Peruvian cohort", *Neuropsychopharmacology*, 45(10), pp. 1595–1602. doi:10.1038/s41386-020-0603-5.

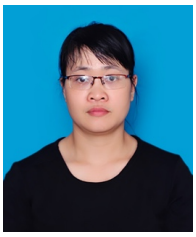
[64] Cardona Tobar, K.M. et al. (2020) "Genome-wide association studies in sheep from Latin America. Review", *Revista mexicana de ciencias pecuarias*, 11(3), pp. 859–883. doi:10.22319/rmcp.v11i3.5372.

[65] Yang, Z. et al. (2021) "Genome-wide association study reveals genetic variations associated with ocean acidification resilience in Yesso scallop Patinopecten yessoensis", *Aquatic Toxicology*, 240, p. 105963. doi:10.1016/j.aquatox.2021.105963.

[66] Peng, W. et al. (2021) "Identification of growth-related SNP and genes in the genome of the Pacific abalone (Haliotis discus hannai) using GWAS", *Aquaculture*, 541, p. 736820. doi:10.1016/j.aquaculture.2021.736820.
[67] Gibbs, R.A. et al. (2003) "The International HapMap Project", *Nature* [Preprint]. doi:10.1038/nature02168.
[68] Siva, N. (2008) "1000 genomes project", *Nature Biotechnology*, 26(3), pp. 256–257.
[69] Jadhav, A., Pramod, D. and Ramanathan, K. (2019) "Comparison of Performance of Data Imputation Methods for Numeric Dataset", *Applied Artificial Intelligence*, 33(10), pp. 913–933. doi:10.1080/08839514.2019.1637138.
[70] Austin, P.C. et al. (2021) "Missing Data in Clinical Research: A Tutorial on Multiple Imputation", *Canadian Journal of Cardiology*, 37(9), pp. 1322–1331. doi:10.1016/j.cjca.2020.11.010.
[71] Choquet, H. et al. (2021) "A large multiethnic GWAS meta-analysis of cataract identifies new risk loci and sex-specific effects", *Nature Communications*, 12(1), p. 3595. doi:10.1038/s41467-021-23873-8.
[72] Powell, V. et al. (2021) "Investigating regions of shared genetic variation in attention deficit/hyperactivity disorder and major depressive disorder: a GWAS meta-analysis", *Scientific Reports*, 11(1), p. 7353. doi:10.1038/s41598-021-86802-1.
[73] Levey, D.F. et al. (2020) GWAS of Depression Phenotypes in the Million Veteran Program and Meta-analysis in More than 1.2 Million Participants Yields 178 Independent Risk Loci. *medRxiv*, p. 2020.05.18.20100685. doi:10.1101/2020.05.18.20100685.
[74] Taherkhani, L. et al. (2022) "The Candidate Chromosomal Regions Responsible for Milk Yield of Cow: A GWAS Meta-Analysis", *Animals*, 12(5), p. 582. doi:10.3390/ani12050582.
[75] Li, J.H. et al. (2021) "Low-pass sequencing increases the power of GWAS and decreases measurement error of polygenic risk scores compared to genotyping arrays", *Genome Research*, 31(4), pp. 529–537. doi:10.1101/gr.266486.120.
[76] Huang, J. et al. (2022) "A Next Generation Sequencing-Based Protocol for Screening of Variants of Concern in Autism Spectrum Disorder", *Cells*, 11(1), p. 10. doi:10.3390/cells11010010.
[77] Chatterjee, N., Shi, J. and García-Closas, M. (2016) "Developing and evaluating polygenic risk prediction models for stratified disease prevention", *Nature Reviews Genetics*, 17(7), pp. 392–406. doi:10.1038/nrg.2016.27.
[78] Natarajan, P. (2018) "Polygenic Risk Scoring for Coronary Heart Disease", *Journal of the American College of Cardiology*, 72(16), pp. 1894–1897. doi:10.1016/j.jacc.2018.08.1041.
[79] Zhang, X. et al. (2018) "Addition of a polygenic risk score, mammographic density, and endogenous hormones to existing breast cancer risk prediction models: A nested case–control study", *PLOS Medicine*, 15(9), p. e1002644. doi:10.1371/journal.pmed.1002644.
[80] Hung, R.J. et al. (2021) "Assessing Lung Cancer Absolute Risk Trajectory Based on a Polygenic Risk Model", *Cancer Research*, 81(6), pp. 1607–1615. doi:10.1158/0008-5472.CAN-20-1237.
[81] Carr, P.R. et al. (2020) "Estimation of Absolute Risk of Colorectal Cancer Based on Healthy Lifestyle, Genetic Risk, and Colonoscopy Status in a Population-Based Study", *Gastroenterology*, 159(1), pp. 129-138.e9. doi:10.1053/j.gastro.2020.03.016.
[82] Darst, B.F. et al. (2021) "Combined Effect of a Polygenic Risk Score and Rare Genetic Variants on Prostate Cancer Risk", *European Urology*, 80(2), pp. 134–138. doi:10.1016/j.eururo.2021.04.013.

SƠ LƯỢC VỀ TÁC GIẢ

Trịnh Thị Xuân



Tốt nghiệp Thạc sỹ ngành Công nghệ Thông tin năm 2007 tại Trường đại học Công nghệ, Đại học Quốc Gia Hà Nội.
Hiện là giảng viên, phó bộ môn cơ sở, khoa công nghệ thông tin, Trường đại học Mở Hà Nội.
Lĩnh vực nghiên cứu: khai phá dữ liệu, tin sinh học.
Email: trinhxuan@hou.edu.vn

Tạ Văn Nhân

Nhận bằng thạc sĩ ngành Khoa học dữ liệu năm 2021, trường Đại học Khoa học tự nhiên, Đại học Quốc gia Hà Nội.
Hiện là chuyên viên tin sinh học tại công ty TNHH LOBI Việt Nam.
Lĩnh vực nghiên cứu: giải trình tự hệ gen người, dự đoán nguy cơ mắc bệnh đa di truyền, dược học hệ gen, mô hình Markov ẩn và mạng Bayes trong tin sinh học.
Email: nhanta@lobi.vn

Hoàng Đỗ Thanh Tùng

Tốt nghiệp Tiến sĩ ngành Khoa học Máy tính tại trường đại học quốc gia Chungbuk, Hàn Quốc.
Hiện là Trưởng phòng Nghiên cứu hệ thống và quản lý, viện Công nghệ Thông tin, VAST.
Lĩnh vực nghiên cứu: các giải pháp lưu trữ, nén, biến đổi, đánh chỉ số và các mô hình cơ sở dữ liệu sinh tinh học nhằm tăng hiệu quả lưu trữ, khai thác thông tin sinh học như về nguồn gốc giống loài, bệnh, thuốc v.v..
Email: tunghdt@gmail.com

Trương Nam Hải

Năm 1988 nhận bằng Tiến sĩ Sinh học phân tử tại Liên Xô. Năm 2004 được phong PGS ngành Sinh học. Năm 2012 được phong GS ngành Sinh học. Nguyên Viện trưởng Viện Công nghệ sinh học và là NCV Cao cấp.
Hiện công tác tại phòng Kỹ thuật di truyền, viện Công nghệ Sinh học, VAST.
Lĩnh vực nghiên cứu: nghiên cứu tạo các protein tái tổ hợp sử dụng trong y dược, nghiên cứu tạo các bộ sinh phẩm chẩn đoán bệnh cho người, động vật, nghiên cứu tạo vaccine bằng protein tái tổ hợp, nghiên cứu gen, loài trong quần thể/khu hệ vi sinh vật bằng kỹ thuật metagenomics.
Email: tnhai@ibt.ac.vn

Trần Đăng Hưng

Tốt nghiệp Thạc sỹ ngành Khoa học máy tính tại trường Đại học Công nghệ, ĐHQGHN năm 2006. Tốt nghiệp Tiến sỹ ngành khai phá tri thức năm 2009 tại viện Khoa học và Công nghệ tiên tiến Nhật Bản (JAIST).
Hiện là Trưởng khoa khoa Công nghệ Thông tin, Trường Đại học Sư phạm Hà Nội.
Lĩnh vực nghiên cứu: khai phá dữ liệu sinh học phân tử, khai phá mạng sinh học, học máy.
Email: hungdt@hnue.edu.vn