ỨNG DỤNG MÔ HÌNH HỌC MÁY PHÂN TÍCH ĐA HÌNH GEN HUYẾT KHỐI LIÊN QUAN ĐẾN SỰ SẢY THAI LIÊN TIẾP

Đặng Huy Hiếu[1], Vũ Quốc Trung[1], Vũ Nguyễn Hồng Phong[1]
Nguyễn Ngọc Thơ[1], Trần Phạm Mạnh[1], Nguyễn Minh Hiền[1]
và Nguyễn Thị Trang[1,2,✉]

[1]Trường Đại Học Y Hà Nội
[2]Bệnh viện Đại học Y Hà Nội

Sảy thai liên tiếp là một biến chứng sản khoa nguy hiểm có liên quan đến các đột biến gen huyết khối. Tuy nhiên, phân tích thủ công các tổ hợp gen là không khả thi với số lượng gen lớn. Do đó, chúng tôi thực hiện nghiên cứu sử dụng mô hình học máy MDR (Multifactor Dimensionality Reduction) nhằm phân tích dữ liệu di truyền và xác định các tổ hợp gen liên quan đến nguy cơ sảy thai liên tiếp trên đối tượng phụ nữ có tiền sử sảy thai. Đây là một nghiên cứu cắt ngang hồi cứu được tiến hành tại Bộ môn Y Sinh học - Di truyền, Trường Đại học Y Hà Nội và Viện Công nghệ DNA và Phân tích Di truyền Genlab. Kết quả phân tích tổ hợp 9 gen cho thấy đối tượng có đồng hợp tử đột biến gen SERPINE1-675 4G/5G và dị hợp tử MTR c.2756A>G (p.Asp919Gly) có nguy cơ sảy thai liên tiếp cao gấp 9,36 lần. Tổ hợp hai gen trên đạt độ chính xác dự đoán là 61,31% với độ nhất quán 8/10 khi kiểm định chéo. Nghiên cứu cho thấy tiềm năng MDR trong việc dự đoán các tổ hợp gen có khả năng tăng nguy cơ sảy thai liên tiếp , từ đó đưa ra những tư vấn và điều trị phù hợp.

Từ khóa: MDR, sảy thai liên tiếp, gen huyết khối, *SERPINE1-675 4G/5G*, *MTR c.2756A>G*.

I. ĐẶT VẤN ĐỀ

Sảy thai liên tiếp (Recurrent pregnancy loss) là biến chứng sản khoa nguy hiểm ảnh hưởng 1 - 4% phụ nữ đang trong độ tuổi sinh sản. Những nguyên nhân được đề cập, bao gồm bất thường nhiễm sắc thể, bất thường tử cung, rối loạn nội tiết, miễn dịch và huyết khối di truyền tuy nhiên vẫn còn khoảng 50% phụ nữ mắc sảy thai liên tiếp chưa xác định được nguyên nhân cụ thể.[1] Những năm gần đây các yếu tố di truyền liên quan đến rối loạn đông máu ngày càng được quan tâm, trong đó, vai trò của các đột biến gen liên quan đến quá trình đông máu như *F5* c.1691G>A (p.Arg506Gln), *F2 c.*97G>A, MTHFR, SERPINE1*... đã được nghiên cứu rộng rãi.[2,3] Tuy nhiên, phần lớn các nghiên cứu hiện nay mới chỉ tập trung vào việc phân tích ảnh hưởng của từng đa hình gen một cách độc lập. Việc tính toán thủ công để tìm ra các tổ hợp đa hình gen là rất khó khăn và dễ bỏ sót những tương tác tiềm ẩn giữa các gen. Sự phát triển của các thuật toán học máy, như MDR (Multifactor Dimensionality Reduction), đã mở ra khả năng phân tích hiệu quả các tổ hợp đa hình gen từ những cơ sở dữ liệu lớn. Để giải quyết hạn chế đó, mô hình Multifactor Dimensionality Reduction (MDR) được phát triển nhằm phân tích tổ hợp gen-gen có liên quan đến kiểu hình bệnh. Mô hình này không chỉ phân loại các tổ hợp theo nguy cơ mà còn có khả năng phát hiện vai trò tiềm ẩn của các gen không có ý nghĩa khi phân tích đơn lẻ, từ đó hỗ trợ tiên lượng nguy cơ sảy thai liên tiếp hiệu quả hơn.

Tác giả liên hệ: Nguyễn Thị Trang
Trường Đại học Y Hà Nội
Email: trangnguyen@hmu.edu.vn
Ngày nhận: 16/07/2025
Ngày được chấp nhận: 30/08/2025

Tại Việt Nam, MDR là hướng nghiên cứu mới trong việc phân tích tương tác giữa các đa hình gen lên biểu hiện bệnh. Tuy nhiên, hiện tại chưa có nghiên cứu nào ứng dụng MDR để phân tích mối tương quan giữa các đa hình gen huyết khối với tình trạng sảy thai liên tiếp . Vì vậy, chúng tôi thực hiện nghiên cứu này với mục tiêu: Ứng dụng mô hình MDR xác định tổ hợp gen có giá trị tiên lượng tình trạng sảy thai thái diễn trên quần thể phụ nữ Việt Nam.

II. ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP

1. Đối tượng

Tiêu chuẩn lựa chọn:

- Người bệnh nữ trong độ tuổi sinh sản, đã có tiền sử sảy thai. Các người bệnh được phân loại thành 2 nhóm: Nhóm sảy thai 1 lần và nhóm người bệnh sảy thai liên tiếp (có ít nhất 2 hoặc nhiều lần sảy thai liên tiếp.[4]

- Người bệnh được thực hiện cận lâm sàng sàng lọc nguyên nhân sảy thai: nhiễm sắc thể đồ; miễn dịch (kháng đông lupus, kháng thể kháng cardiolipin, kháng beta-2 glycoprotein I); siêu âm vùng chậu, chụp X-quang và nội soi buồng tử cung; xét nghiệm nội tiết - chuyển hoá (TSH, prolactin, dung nạp glucose nếu chỉ định); cùng các xét nghiệm khác (công thức máu, FSH, estradiol, progesterone…)- Người bệnh được xét nghiệm gen bằng PCR theo bộ xét nghiệm gen huyết khối tại Viện Công nghệ DNA và phân tích di truyền Genlab gồm 9 gen: *F2* c.*97G>A*, F5* c.1691G>A (p.Arg506Gln)*, MTHFR* c.677C>T (p.Ala222Val)*, MTHFR* c.1298A>C (p.Glu433Ala)*, SERPINE1 -675 4G/5G, F13A1* c.103G>T (p.Val34Leu)*, FGB* β-455 G>A*, MTR* c.2756A>G (p.Asp919Gly)*, MTRR* c.66A>G (p.Ile22Met).

Tiêu chuẩn loại trừ:

- Người bệnh có thai tống xuất ra khỏi buồng tử cung sau 22 tuần hoặc trọng lượng trên 500 gram.

- Người bệnh đình chỉ thai chủ động/phá thai theo chỉ định của bác sĩ.

- Người bệnh sảy thai do các nguyên nhân đã xác định[5]:

+ Bất thường cấu trúc giải phẫu tử cung.

+ Bất thường nhiễm sắc thể thai nhi.

+ Bệnh lý nội tiết: thiếu hụt pha hoàng thể, rối loạn androgen (bao gồm rối loạn LH), rối loạn prolactin, đái tháo đường, rối loạn chức năng tuyến giáp.

+ Hội chứng kháng phospholipid, các bệnh lý tự miễn hệ thống khác.

- Hồ sơ người bệnh không có đầy đủ thông tin.

2. Phương pháp

Thiết kế nghiên cứu

Nghiên cứu cắt ngang hồi cứu từ tháng 10/2019 đến tháng 10/2023 tại Bộ môn Y Sinh học - Di truyền, Trường Đại học Y Hà Nội và Viện công nghệ DNA và phân tích di truyền Genlab.

Phương pháp chọn mẫu

Chọn mẫu thuận tiện. Nghiên cứu tuyển chọn được 826 người bệnh thỏa mãn tiêu chuẩn chọn mẫu, bao gồm 176 người bệnh sảy thai 1 lần và 176 người bệnh sảy thai liên tiếp .

Quy trình nghiên cứu:

- Bước 1: Hồi cứu lại hồ sơ bệnh án và kết quả xét nghiệm tại Viện công nghệ DNA và phân tích di truyền Genlab, lựa chọn đối tượng theo tiêu chuẩn lựa chọn và tiêu chuẩn loại trừ.

- Bước 2: Mã hóa ẩn danh cho bộ dữ liệu thu thập được. Các người bệnh phù hợp tham gia nghiên cứu được loại bỏ các thông tin cá nhân có thể định danh, chỉ giữ lại kết quả xét nghiệm gen và kiểu hình sảy thai. Mỗi người bệnh được mã hóa bằng 1 mã nghiên cứu riêng ứng với bộ dữ liệu gốc ban đầu.

- Bước 3: Phân tích đặc điểm đa hình

đơn gen của 9 gen huyết khối liên quan đến quá trình tăng đông ở phụ nữ: *F2* c.*97G>A, *F5* c.1691G>A (p.Arg506Gln), *MTHFR* c.677C>T (p.Ala222Val), *MTHFR* c.1298A>C (p.Glu433Ala), *SERPINE1* -675 4G/5G, *F13A1* c.103G>T (p.Val34Leu), *FGB* β-455 G>A, *MTR* c.2756A>G (p.Asp919Gly), *MTRR* c.66A>G (p.Ile22Met).

- Bước 4: Sử dụng mô hình học máy MDR tìm ra tổ hợp gen có ý nghĩa tiên lượng với tình trạng sảy thai liên tiếp.

Xử lý số liệu

Tổng quan về phần mềm MDR

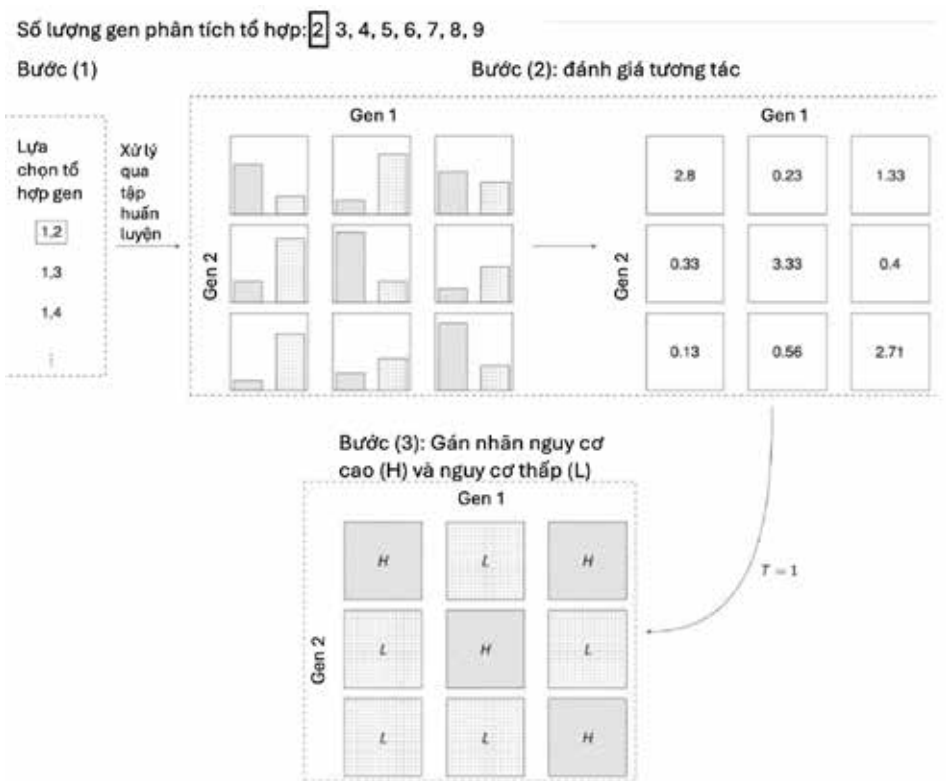
- Nghiên cứu sử dụng phần mềm MDR phiên bản 3.0.2 (https://mybiosoftware.com/mdr-2-0-multifactor-dimensionality-reduction.html#google_vignette).

- Nguyên tắc hoạt động của thuật toán MDR:

MDR (Multifactor Dimensionality Reduction) là công cụ học máy giúp đánh giá tương tác giữa các gen với kiểu hình cụ thể, khắc phục được 2 nhược điểm chính trong các nghiên cứu phân tích gen: khó khăn khi phân tích đa biến bằng các phương pháp thống kê truyền thống, bộ dữ liệu nhỏ không có ý nghĩa thống kê.

MDR tổng hợp tất cả các tổ hợp gen có thể xảy ra từ số lượng biến được lựa chọn. Ví dụ: Với 3 gen, mỗi gen có 3 kiểu gen tổng hợp được 3^3 = 27 tổ hợp có khả năng. Tỷ lệ kiểu hình nhóm mang bệnh và nhóm đối chứng của các tổ hợp được thống kê theo bộ số liệu.

Các tổ hợp được gán nhãn là nguy cơ cao nếu tỷ lệ ca bệnh/ca đối chứng > 1 và nguy cơ thấp nếu tỷ lệ này < 1. Sau khi phân loại được nguy cơ, MDR sẽ chuyển từ phân tích đa biến thành phân tích đơn biến (có/không có tổ hợp gen và mang bệnh/không mang bệnh).



Hình 1. Nguyên tắc hoạt động của thuật toán MDR

- Thuật toán MDR sử dụng phương pháp kiểm định chéo (Cross-validation) khắc phục nhược điểm của bộ số liệu nhỏ. Tập dữ liệu sẽ được chia thành nhiều tập con, trong đó một tập dữ liệu con sử dụng để kiểm định và các tập dữ liệu còn lại dùng để huấn luyện thuật toán. Các tập con sẽ được chia ngẫu nhiên và quá trình huấn luyện, kiểm định sẽ được lặp lại nhiều lần. Các tổ hợp gen có ý nghĩa qua các lần chạy số liệu được đánh giá bằng các tiêu chí sau:

Tính nhất quán khi kiểm định chéo (Cross-Validation Consistency): được định nghĩa là tổng số lần có tổ hợp gen trùng khớp nhau trên tổng số lần kiểm định dữ liệu. Tổ hợp gen xuất hiện càng nhiều, tính nhất quán càng cao, ý nghĩa của tổ hợp càng lớn.

Độ nhạy và độ đặc hiệu trong tập huấn luyện và tập kiểm định.

Độ chính xác cân bằng: trung bình cộng của độ nhạy và độ đặc hiệu.

- Kết quả trả về là các tổ hợp có giá trị nhất với mô hình 1 gen, 2 gen, 3 gen, cho tới 9 gen.

- Tổ hợp gen có tính nhất quán cao, độ chính xác cân bằng cao được lựa chọn để tiếp tục xử lý. Trong nghiên cứu này, chúng tôi lựa chọn các tổ hợp gen có tỷ lệ > 1% để phân tích sâu hơn.

- Các tham số thống kê được xử lý bằng phương pháp thống kê y học theo phần mềm Stata 16.0, các giá trị p < 0,05 được coi là có ý nghĩa thống kê.

3. Đạo đức nghiên cứu

- Nghiên cứu được thông qua Hội đồng Đạo đức trường Đại học Y Hà Nội số 7818/QĐ-ĐHYHN, tuân thủ chặt chẽ Hướng dẫn quốc gia về đạo đức trong Nghiên cứu y sinh học năm 2013.

- Các bệnh nhân được giải thích mục đích và nội dung nghiên cứu rõ ràng. Quá trình xét nghiệm và thu thập thông tin chỉ được tiến hành khi có sự đồng ý của đối tượng nghiên cứu.

- Các thông tin chỉ được sử dụng cho mục đích của nghiên cứu và không nhằm mục đích khác.

- Đối tượng nghiên cứu có quyền dừng tham gia hoặc rút khỏi nghiên cứu tại bất kì thời điểm nào.

III. KẾT QUẢ

1. Đặc điểm đa hình gen huyết khối của nhóm đối tượng nghiên cứu

Nghiên cứu chúng tôi thực hiện trên quần thể 826 người bệnh, trong đó bao gồm 176 người bệnh sảy thai 1 lần và 650 người bệnh sảy thai liên tiếp. Kết quả phân tích gen cho thấy có sự khác biệt về tỷ lệ các kiểu gen của các gen *SERPINE1* -675 4G/5G, *FGB β-455 G>A* (p < 0,0001) và *MTRR* c.66A>G (p.Ile22Met) (p=0,02) giữa 2 nhóm sảy thai 1 lần và sảy thai liên tiếp . Không có khác biệt có ý nghĩa thống kê giữa 2 nhóm ở các đa hình gen còn lại (Bảng 1).

Bảng 1. Đặc điểm đa hình gen huyết khối ở 2 nhóm sảy thai 1 lần và sảy thai liên tiếp

Gen	Sảy thai 1 lần (n = 176)			Sảy thai liên tiếp (n = 650)			p-value
	*AA** (n, %)	*Aa* (n, %)	*aa* (n, %)	*AA** (n, %)	*Aa* (n, %)	*aa* (n, %)	
***F2* c.*97G>A**	173 (98,3)	3 (1,7)	0 (0)	633 (97,38)	15 (2,31)	2 (0,31)	0,862

Gen	Sảy thai 1 lần (n = 176)			Sảy thai liên tiếp (n = 650)			p-value
	*AA** (n, %)	*Aa* (n, %)	*aa* (n, %)	*AA** (n, %)	*Aa* (n, %)	*aa* (n, %)	
***F5* c.1691G>A (p.Arg506Gln)**	176 (100)	0	0	642 (98,77)	8 (1,23)	0	0,214
***MTHFR* c.677C>T (p.Ala222Val)**	98 (55,68)	63 (35,80)	15 (8,52)	366 (56,31)	237 (36,46)	47 (7,23)	0,846
***MTHFR* c.1298A>C (p.Glu433Ala)**	95 (53,98)	67 (38,07)	14 (7,95)	336 (51,69)	256 (39,38)	58 (8,92)	0,840
***SERPINE1* -675 4G/5G**	87 (49,43)	67 (38,07)	22 (12,5)	151 (23,23)	315 (48,46)	184 (28,31)	**0,000**
***F13A1* c.103G>T (p.Val34Leu)**	172 (97,73)	4 (2,27)	0 (0,2)	640 (98,46)	9 (1,38)	1 (0,15)	0,600
***FGB* β-455 G>A**	57 (32,39)	88 (50,0)	31 (17,61)	323 (49,69)	260 (40,00)	67 (10,31)	**0,000**
***MTR* c.2756A>G (p.Asp919Gly)**	117 (66,48)	45 (25,57)	14 (10,4)	479 (73,69)	136 (20,92)	35 (5,38)	0,142
***MTRR* c.66A>G (p.Ile22Met)**	105 (59,66)	58 (32,95)	13 (7,39)	311 (47,85)	282 (43,38)	57 (8,77)	**0,020**

**Quy ước AA: đồng hợp tử bình thường, Aa: dị hợp tử, aa: đồng hợp tử đột biến*

2. Kết quả phân tích tổ hợp đa hình gen huyết khối bằng phần mềm MDR

Kết quả cho thấy gen *SERPINE1* -675 4G/5G có giá trị dự báo sảy thai liên tiếp cao nhất. Bảng 2 minh hoạ các tổ hợp được phân tích cùng gen *SERPINE1* -675 4G/5G. Tổ hợp 2 gen *SERPINE1* -675 4G/5G, *MTR* c.2756A>G (p.Asp919Gly) có độ chính xác cao nhất trong các tổ hợp (61.31%) với độ nhất quán 8/10. Tổ hợp 7 gen *SERPINE1* -675 4G/5G*, MTR* c.2756A>G (p.Asp919Gly)*, MTHFR* c.677C>T (p.Ala222Val)*, MTHFR* c.1298A>C (p.Glu433Ala)*, FGB* β-455 G>A*, MTRR* c.66A>G (p.Ile22Met*), F13A1* c.103G>T (p.Val34Leu) có độ nhất quán 10/10 mặc dù độ chính xác tập kiểm định ở mức chấp nhận được (>51.09%). Các tổ hợp còn lại mặc dù có độ nhất quán 10/10 nhưng độ chính xác trên tập kiểm định thấp do hiện tượng quá khớp (overfitting) của thuật toán nên không được lựa chọn để phân tích sâu hơn.

Bảng 2. Kết quả phân tích MDR tương tác gen

Mô hình đột biến gen huyết khối	*SERPINE1* -675 4G/5G						
	MTR c.2756A>G (p.Asp919Gly)						
	MTHFR c.677C>T (p.Ala222Val)						
	MTHFR c.1298A>C (p.Glu433Ala)						
	FGB β-455 G>A						
	MTRR c66A>G (p.Ile22Met)						
	F13A1 c.103G>T (p.Val34Leu)						
	F2 c.*97G>A						
	F5 c.1691G>A (p.Arg506Gln)						
Độ chính xác cân bằng tập huấn luyện (%)		63,1	63,79	77,96	78,45	78,54	78,54
Độ chính xác cân bằng tập kiểm định (%)		63,1	61,34	50,55	51,09	50,31	50,31
Độ nhạy (%)		76,77	71,54	69,85	70,92	71,08	71,08
Độ đặc hiệu (%)		49,43	51,14	31,25	31,25	29,55	29,55
Độ nhất quán (CVC)		10/10	8/10	10/10	10/10	10/10	10/10
p value mô hình		< 0,0001	< 0,0001	< 0,0001	< 0,0001	< 0,0001	< 0,0001

Với tổ hợp 2 gen *SERPINE1* -675 4G/5G và *MTR* c.2756A>G (p.Asp919Gly), có $3^2 = 9$ tổ hợp có thể xảy ra và đều xuất hiện trong nghiên cứu của chúng tôi, trong đó 4 tổ hợp có yếu tố nguy cơ cao. 8 tổ hợp có tần số > 1% được phân tích mối liên hệ với tình trạng sảy thai liên tiếp . Kết quả cho thấy kiểu gen đồng hợp của *SERPINE1* -675 4G/5G và dị hợp của *MTR* c.2756A>G (p.Asp919Gly) có nguy cơ sảy thai cao, với OR = 9,36, p = 0,004 (Bảng 3).

Bảng 3. Phân tích các tổ hợp 2 gen

Tổ hợp 2 gen		Sảy liên tiếp	Sảy 1 lần	p value	OR	95% CI
SERPINE1 -675 4G/5G	*MTR* c.2756A>G (p.Asp919Gly)					
Aa	*AA*	234	48	0,0303	1,5	1,025 - 2,217
aa	*AA*	149	19	0,0004	2458	1,459 - 4,332
AA	*AA*	96	50	< 0,0001	0,437	0,290 - 0,662
aa	*Aa*	33	1	0,004	**9,360**	1,541 - 382,8
AA	*Aa*	42	31	< 0,0001	0,323	0,191 - 0,552

**Quy ước AA: đồng hợp tử bình thường; Aa: dị hợp tử; aa: đồng hợp tử đột biến*

Với tổ hợp 7 gen *SERPINE1* -675 4G/5G, *MTR* c.2756A>G (p.Asp919Gly), *MTHFR* c.677C>T (p.Ala222Val), *MTHFR* c.1298A>C (p.Glu433Ala), *MTRR* c.66A>G (p.Ile22Met), *F13A1* c.103G>T (p.Val34Leu), *FGB* β-455 G>A có 3^7 = 2187 tổ hợp có thể xảy ra và 264 xuất hiện trong nghiên cứu của chúng tôi, trong đó 173 tổ hợp có yếu tố nguy cơ cao. 23 tổ hợp có tần số >1% được phân tích mối liên hệ với tình trạng sảy thai liên tiếp (Bảng 4). Kết quả cho thấy có 1 tổ hợp 7 gen có ý nghĩa với OR = 0,392, p = 0,021.

Bảng 4. Phân tích tổ hợp 7 gen

Tổ hợp 7 gen		Sảy liên tiếp	Sảy 1 lần	p value	OR	95% CI
SERPINE1 -675 4G/5G	*Aa*	15	10	0,021	0,392	0,162 - 0,964
MTR c.2756A>G (p.Asp919Gly)	AA					
MTHFR c.677C>T (p.Ala222Val)	AA					
MTHFR c.1298A>C (p.Glu433Ala)	Aa					
FGB β-455 G>A	Aa					
*MTRR c.*66A>G (p.Ile22Met)	AA					
F13A1 c.103G>T (p.Val34Leu)	AA					

**Quy ước AA: đồng hợp tử bình thường, Aa: dị hợp tử, aa: đồng hợp tử đột biến*

IV. BÀN LUẬN

Sảy thai liên tiếp là biến chứng sản khoa thường gặp do rối loạn đông máu, nội tiết hay các bất thường về miễn dịch, giải phẫu tử cung. Nghiên cứu của Yuanjja Wen cho thấy gần 5% phụ nữ sẽ bị sảy thai từ 2 lần trở nên trong đời, đặt ra nhu cầu xác định các nguyên nhân cụ thể gây ra tình trạng này.[1] Một trong các hướng nghiên cứu được quan tâm là phân tích ảnh hưởng của các đột biến gen huyết khối lên tình trạng sảy thai liên tiếp. Tuy nhiên, số lượng đột biến gen huyết khối được phát hiện ngày càng lớn, dẫn đến khó khăn khi phân tích ảnh hưởng của nhiều gen bằng các phương pháp thống kê thông thường. Do đó, thuật toán MDR là một giải pháp tiềm năng để hỗ trợ các nhà nghiên cứu phân tích tương tác giữa các đột biến gen huyết khối với nguy cơ sảy thai liên tiếp.

Trong nghiên cứu này, chúng tôi thực hiện trên 9 đột biến gen phổ biến trên quần thể phụ nữ có tiền sử sảy thai ở Việt Nam (*F2* c.*97G>A*, F5* c.1691G>A (p.Arg506Gln)*, MTHFR* c.677C>T (p.Ala222Val)*, MTHFR* c.1298A>C (p.Glu433Ala)*, SERPINE1* -675 4G/5G*, F13A1* c.103G>T (p.Val34Leu)*, FGB* β-455 G>A *, MTR* c.2756A>G (p.Asp919Gly)*, MTRR* c.66A>G (p.Ile22Met)*.* Kết quả xử lý số liệu cho thấy có chênh lệch đáng kể giữa nhóm sảy thai liên tiếp và nhóm sảy thai một lần ($p < 0,05$) ở các đột biến gen phổ biến, bao gồm *SERPINE1* -675 4G/5G (76,77% với 50,57%), *MTRR* c.66A>G (p.Ile22Met) (52,15% với 40,34%) và *FGB* β-455 G>A (50,31% với 67,31%). Kết quả nghiên cứu trước đây cho thấy các đột biến gen này xuất hiện với tần số lớn và có ý nghĩa tiên lượng với tình trạng sảy thai liên tiếp (OR *SERPINE1* -675 4G/5G: 1,67; *MTRR* c.66A>G (p.Ile22Met): 1,22; *FGB* β-455 G>A: 1,6).[1,6]

Nghiên cứu cũng ghi nhận tỷ lệ đột biến rất thấp ở cả nhóm sảy thai liên tiếp và sảy thai 1 lần ở các gen *F2* c.*97G>A và *F5* c.1691G>A (p.Arg506Gln), tuy nhiên khác biệt không có ý nghĩa thống kê. Theo Jaskamal và cộng sự, tỷ lệ đột biến *F5* c.1691G>A (p.Arg506Gln) và *F2* c.*97G>A dao động trong khoảng 3-15% ở người châu Âu và có thể thấp hơn ở các chủng tộc khác.[7] Mặc dù vậy, phân tích gộp từ 89 nghiên cứu khẳng định 2 đột biến này làm tăng nguy cơ sảy thai liên tiếp (OR *F5* c.1691G>A (p.Arg506Gln): 2,44; *F2* c.*97G>A: 2,08[8]).

Nghiên cứu chúng tôi sử dụng thuật toán MDR và phân tích các tổ hợp gen có ý nghĩa tiên lượng với từng mô hình. Kết quả trả về là các mô hình tiên lượng nguy cơ sảy thai liên tiếp khi đã từng sảy thai 1 lần. Thực tế, các nghiên cứu trước đây tập trung so sánh với quần thể phụ nữ khỏe mạnh để tìm ra đột biến gen nguy cơ, từ đó gợi ý các đột biến gen cần sàng lọc trước khi mang thai.[1] liên tiếp Tuy nhiên, theo Harvey và cộng sự, nguy cơ sảy thai liên tiếp tăng đáng kể sau mỗi lần sảy thai.[9] Do đó, hướng tiếp cận này giúp tìm ra các tổ hợp gen mang ý nghĩa tiên lượng cho các lần mang thai tiếp theo ở phụ nũ có tiền sử sảy thai không xác định được nguyên nhân, từ đó có kế hoạch theo dõi và điều trị phù hợp.

Ưu điểm của thuật toán MDR khi phân tích tương tác đa gen là chuyển đổi từ đa biến thành đơn biến, giúp đơn giản hóa quá trình xử lý số liệu. Với khả năng kiểm định chéo, MDR khắc phục được nhược điểm của bộ số liệu nhỏ, không có ý nghĩa thống kê. Đặc biệt, MDR có thể tìm ra các đột biến gen không có ý nghĩa khi phân tích đơn lẻ, nhưng lại có giá trị tiên lượng lớn trong tổ hợp đa hình gen.[10] Nghiên cứu của Trifonova đã sử dụng thuật toán MDR phân tích tương tác gen huyết khối với kiểu hình sảy thai liên tiếp . Kết quả cho thấy đa hình đồng hợp tử D/D của gen *ACE* và 4G/4G của gen *SERPINE-1* khi phân tích riêng lẻ không

thấy sự khác biệt có ý nghĩa thống kê nhưng khi phân tích tổng hợp 2 gen lại cho thấy nguy cơ phát triển sảy thai liên tiếp trước tuần thứ 25 của thai kỳ tăng lên đáng kể ở phụ nữ châu Âu.[11] Điều này cho thấy ứng dụng tiềm năng của MDR cho các phân tích đa gen.

Kết quả chúng tôi cho thấy, mô hình 2 gen *SERPINE1* -675 4G/5G và *MTR* c.2756A>G (p.Asp919Gly) có ý nghĩa tiên lượng, với độ chính xác 61,34%, độ nhất quán 8/10 với p value mô hình < 0,0001. Trong mô hình đơn gen, *SERPINE1* -675 4G/5G có giá trị tiên lượng cao nhất trong bộ số liệu với độ nhất quán 10/10 và độ chính xác 61,3%. Đây là đột biến có sự khác biệt lớn nhất giữa 2 nhóm bệnh và nhóm chứng (76,77% với 50,57%, p < 0,05). Đáng chú ý, đột biến *MTR* c.2756A>G (p.Asp919Gly) mặc dù có tần số xuất hiện cao, nhưng không có khác biệt giữa 2 nhóm khi phân tích đơn lẻ (36% với 26,3%, p = 0,142). Bên cạnh đó, *FGB* β-455 G>A và *MTRR* c.66A>G (p.Ile22Met) lại không xuất hiện trong mô hình tiên lượng 2 gen, mặc dù 2 đột biến này có khác biệt đáng kể giữa 2 nhóm khi phân tích đơn lẻ.

Phân tích sâu hơn mô hình 2 gen *SERPINE1* -675 4G/5G và *MTR* c.2756A>G (p.Asp919Gly) cho thấy sự kết hợp của 2 đột biến này làm tăng đáng kể nguy cơ sảy thai liên tiếp . Đặc biệt, tổ hợp đồng hợp tử đột biến của *SERPINE1* -675 4G/5G với dị hợp tử đột biến của *MTR* c.2756A>G (p.Asp919Gly) làm tăng nguy cơ gấp 9,36 lần (OR = 9,36, p = 0,004). Các nghiên cứu tại Việt Nam cho thấy tỷ lệ đột biến ở nhóm sảy thai liên tiếp với gen *SERPINE1* -675 4G/5G là 79,41%, với gen *MTR* c.2756A>G (p.Asp919Gly) là 39,13%, nhưng không liên quan đến sảy thai liên tiếp, mặc dù các bằng chứng trước đây đã khẳng định nguy cơ từ 2 đột biến này.[1,12,13] Chúng tôi đề xuất các nghiên cứu sâu hơn về ảnh hưởng của sự kết hợp 2 đột biến gen này lên tình trạng sảy thai liên tiếp trên quần thể Việt Nam.

Nghiên cứu cũng cho thấy tổ hợp 7 gen *SERPINE1* -675 4G/5G, *MTR* c.2756A>G (p.Asp919Gly)*, MTHFR*c.677C>T(p.Ala222Val)*, MTHFR* c.1298A>C (p.Glu433Ala)*, FGB* β-455 G>A*, MTRR* c.66A>G (p.Ile22Met) và *F13A1* c.103G>T (p.Val34Leu) có ý nghĩa tiên lượng. Khi phân tích sâu hơn, chúng tôi tìm ra được tổ hợp gen làm giảm nguy cơ sảy thai liên tiếp (OR = 0,392, p = 0,021). Mặc dù, mô hình tổ hợp 7 gen có độ nhất quản cao 10/10 nhưng độ chính xác chỉ dừng lại ở mức chấp nhận được (51,09%), cho thấy đây không phải là mô hình đáng tin cậy. Nguyên nhân lý giải cho điều nay là do khi số lượng gen trong tổ hợp gen tăng lên càng nhiều, nguy cơ xảy ra hiện tượng quá khớp (overfitting) là càng lớn. Overfitting xảy ra khi lựa chọn huấn luyện mô hình với những tổ hợp có quá nhiều biến, khiến cho mô hình chỉ có độ chính xác cao đối với tập huấn luyện, và sai lệch nhiều trên tập kiểm định, từ đó làm giảm ý nghĩa tiên lượng thực tế của mô hình. Do đó, cần có bộ dữ liệu lớn và đa dạng hơn để khắc phục nhược điểm này của thuật toán.[14]

Nghiên cứu của chúng tôi có một số hạn chế. Có sự chênh lệch trong quá trình chọn mẫu nghiên cứu giữa nhóm sảy thai liên tiếp và sảy thai 1 lần, do đó có thể ảnh hưởng đến kết quả nghiên cứu. Việc lựa chọn nhóm chứng sảy thai một lần thay vì nhóm chứng hoàn toàn khoẻ mạnh như các nghiên cứu trước đây gây khó khăn trong việc đối chiếu với các kết quả đã có. Cuối cùng, nghiên cứu chưa đánh giá được liên quan của các biến thể gen với các tình trạng lâm sàng khác như tuổi, triệu chứng lâm sàng do sự thiếu hụt về dữ liệu nghiên cứu, do đó các tổ hợp gen có ý nghĩa cần được nghiên cứu sâu hơn.

Tuy nhiên, đây là nghiên cứu đầu tiên tại Việt Nam sử dụng thuật toán MDR để phân tích ảnh hưởng của các đột biến gen huyết khối lên tình trạng sảy thai liên tiếp , từ đó tìm ra được tổ hợp gen có ý nghĩa tiên lượng hỗ trợ cho

quyết định lâm sàng. Vì vậy, chúng tôi kiến nghị cần tiến hành các nghiên cứu tương đương có cỡ mẫu lớn hơn trong tương lai để có thêm các bằng chứng xác thực về ảnh hưởng của các tổ hợp gen với tình trạng sảy thai liên tiếp cũng nhưng mở rộng ứng dụng mô hình MDR trong phân tích gen với các tình trạng lâm sàng khác.

V. KẾT LUẬN

Nghiên cứu của chúng tôi sử dụng mô hình MDR để phân tích mối liên quan giữa các đột biến gen huyết khối và tình trạng sảy thai liên tiếp . Kết quả ghi nhận mô hình 2 gen *SERPINE1* -675 4G/5G và *MTR* c.2756A>G (p.Asp919Gly) có ý nghĩa tiên lượng, trong đó sự kết hợp giữa đồng hợp tử đột biến *SERPINE1*-675 4G/5G và dị hợp tử đột biến *MTR* c.2756A>G (p.Asp919Gly) làm tăng nguy cơ lên gấp 9,36 lần.

LỜI CẢM ƠN

Chúng tôi xin gửi lời cảm ơn chân thành tới các nhân viên Bộ môn Y Sinh học - Di truyền, Trường Đại học Y Hà Nội và Viện công nghệ DNA đã hỗ trợ chúng tôi thu thập số liệu nghiên cứu này. Chúng tôi cam kết không xung đột lợi ích từ kết quả nghiên cứu này.

TÀI LIỆU THAM KHẢO

1. Wen Y, He H, Zhao K Thrombophilic gene polymorphisms and recurrent pregnancy loss: a systematic review and meta-analysis. *Journal of Assisted Reproduction and Genetics*. 2023 May 30; 40(7)doi:10.1007/s10815-023-02823-x.

2. Factor V Leiden. doi:10.1002/ajh.24222.

3. Li X, Liu Y, Zhang R, Tan J, Chen L, Liu Y Meta-Analysis of the Association between Plasminogen Activator Inhibitor-1 4G/5G Polymorphism and Recurrent Pregnancy Loss. *Medical Science Monitor : International Medical Journal of Experimental and Clinical Research*. 2015 Apr 11; 21doi:10.12659/MSM.892898.

4. Definitions of infertility and recurrent pregnancy loss: a committee opinion - PubMed. *Fertility and sterility*. 2013 Jan; 99(1) doi:10.1016/j.fertnstert.2012.09.023.

5. Alijotas-Reig J, Garrido-Gimenez C Current concepts and new trends in the diagnosis and management of recurrent miscarriage - PubMed. *Obstetrical & gynecological survey*. 2013 Jun; 68(6)doi:10.1097/OGX.0b013e31828aca19.

6. Shaker MM, Elaraby NM, Shalabi TA, et al Association of MTR and MTRR polymorphisms with recurrent pregnancy loss: a case control study. *Molecular Biology Reports. 2024 51:1*. 2024-09-09; 51(1)doi: 10.1007/s11033-024-09860-4.

7. Padda J, Khalid K, Mohan A et al Factor V Leiden G1691A and Prothrombin Gene G20210A Mutations on Pregnancy Outcome - PubMed. *Cureus*. 08/15/2021; 13(8) doi:10.7759/cureus.17185.

8. Liu X, Chen Y, Ye C, Xing D, et al Hereditary thrombophilia and recurrent pregnancy loss: a systematic review and meta-analysis - PubMed. *Human reproduction (Oxford, England)*. 04/20/2021; 36(5)doi:10.1093/humrep/deab010.

9. Risch HA, Weiss NS, Aileen Clarke E, Miller AB Risk factors for spontaneous abortion and its recurrence. *American Journal of Epidemiology*. 1988/08/01; 128(2)doi: 10.1093/oxfordjournals.aje.a114982.

10. Hahn LW, Ritchie MD, Moore JH Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions - PubMed. *Bioinformatics (Oxford, England)*. 02/12/2003; 19(3)doi: 10.1093/bioinformatics/btf869.

11. Trifonova EA, Swarovskaya MG, Ganzha OA, et al The interaction effect of angiogenesis and endothelial dysfunction-related gene variants increases the susceptibility of

recurrent pregnancy loss. *Journal of Assisted Reproduction and Genetics. 2019;* 36:4. 2019-01-24; 36(4)doi:10.1007/s10815-019-01403-2.

12. Hoàng Thị Ngọc Lan, Trần Sơn Tùng, Phan Thị Thu Giang et al Nghiên cứu một số biến thể di truyền gây tăng nguy cơ huyết khối ở phụ nữ mất thai tái diễn. *Tạp chí Y học Việt Nam*. 2023/05/16; 526(1A). doi:10.51298/vmj.v526i1A.5306.

13. Trần Ngọc Thảo My, Triệu Tiến Sang, Nguyễn Ngọc Nhất, Trần Văn Tuấn Đặc điểm kiểu gen của đa hình MTHFR C677T, MTHFR A1298C, MTR A2756G và MTRR A66G ở phụ nữ sảy thai liên tiếp. *Bản B của Tạp chí Khoa học và Công nghệ Việt Nam*. 2022/02/25; 64(2) doi: 10.31276/VJST.64(2).01-04.

14. Hahn Lance W, Ritchie Marylyn D, Moore Jason H. Hahn LW, Ritchie MD, Moore JH Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics*. 2003/02/12; 19(3) doi:10.1093/bioinformatics/btf869.

Summary

APPLICATION OF MACHINE LEARNING MODELS TO ANALYZE THROMBOPHILIA-RELATED GENE POLYMORPHISMS IN RECURRENT MISCARRIAGE

Recurrent miscarriage is a serious obstetric complication that has been associated with thrombophilic gene mutations. However, manual analysis of gene combinations is not feasible due to the large number of genes involved. Therefore, we conducted a study using the Multifactor Dimensionality Reduction (MDR) machine learning model to analyze genetic data and identify gene combinations associated with the risk of recurrent miscarriage in women with a history of pregnancy loss. This is a retrospective cross-sectional study conducted at the Department of Biomedical Genetics, Hanoi Medical University, and the Institute of DNA Technology and Genetic Analysis – Genlab. Analysis of nine gene combinations revealed that individuals with homozygous PAI-1 4G/5G mutations and heterozygous MTR A2756G variants had a 9.36 times increased risk of recurrent miscarriage. This two-gene combination yielded a predictive accuracy of 61.31% and a cross-validation consistency of 8 out of 10. The study highlights the potential of MDR in predicting gene combinations associated with an increased risk of recurrent miscarriage, thereby supporting appropriate counseling and treatment strategies.

Keywords: MDR, recurrent miscarriage, thrombophilic gene, *SERPINE1-675* 4G/5G, *MTR c.2756A>G*.