ĐÁNH GIÁ HIỆU NĂNG MÔ HÌNH GAIL ĐIỀU CHỈNH VÀ GAIL-ROSNER-COLDITZ TRONG DỰ BÁO NGUY CƠ UNG THƯ VÚ TẠI VIỆT NAM

Trần Thị Thanh Hương[1,2], Lưu Ngọc Minh[2]
Nguyễn Hương Giang[1] và Bùi Thị Oanh[1,2,✉]
[1]Bệnh viện K
[2]Trường Đại học Y Hà Nội

Nghiên cứu hồi cứu 31.016 phụ nữ tham gia sàng lọc ung thư vú, ghi nhận 19 trường hợp ung thư vú, nhằm đánh giá hiệu năng của mô hình Gail điều chỉnh và mô hình kết hợp Gail-Rosner-Colditz. Dữ liệu được chia ngẫu nhiên bằng phương pháp lấy mẫu phân tầng thành tập huấn luyện 60% và tập kiểm tra 40%; mất cân bằng dữ liệu được xử lý bằng giảm mẫu ngẫu nhiên kết hợp SMOTE. Hiệu năng được đánh giá bằng ROC-AUC, PR-AUC và chỉ số Brier; các yếu tố liên quan với nguy cơ mắc ung thư vú được phân tích bằng hồi quy logistic phạt Firth. Mô hình kết hợp làm tăng ROC-AUC ở cả ba thuật toán; Weighted Kernel Logistic Regression đạt ROC-AUC cao nhất 0,662 ± 0,062. Tuổi có liên quan với nguy cơ mắc ung thư vú, trong khi sinh 2 con và sinh 3 con có liên quan với giảm nguy cơ mắc so với nhóm chưa sinh con. Mô hình kết hợp Gail-Rosner-Colditz cho hiệu năng dự báo cao hơn mô hình Gail điều chỉnh, nhưng giá trị dự báo vẫn còn hạn chế do số biến cố thấp.

Từ khóa: Ung thư vú, mô hình Gail, Rosner-Colditz, dự báo nguy cơ, học máy.

I. ĐẶT VẤN ĐỀ

Ung thư vú là bệnh ung thư thường gặp nhất ở nữ giới trên toàn cầu; tại Việt Nam, đây cũng là ung thư có số ca mới mắc đứng hàng đầu theo ước tính GLOBOCAN 2022.[1] Phân tầng nguy cơ cá thể hóa có vai trò quan trọng trong lựa chọn chiến lược sàng lọc, tư vấn dự phòng và tối ưu hóa phân bổ nguồn lực. Trong số các công cụ đánh giá nguy cơ dựa trên bảng hỏi, mô hình Gail và mô hình Rosner-Colditz là hai mô hình được sử dụng rộng rãi nhất.[2-5] Tuy nhiên, hiệu năng phân biệt của các mô hình này thường chỉ ở mức khiêm tốn và cần được hiệu chỉnh theo từng quần thể cụ thể.[5] Ngoài ra, dữ liệu sàng lọc ung thư vú ngoài thực hành thường có tỷ lệ biến cố rất thấp, làm nảy sinh vấn đề mất cân bằng dữ liệu nghiêm trọng, gây khó khăn cho cả mô hình thống kê truyền thống lẫn các thuật toán học máy.[6-9]

Mô hình Gail được Gail và cộng sự phát triển từ năm 1989 nhằm ước tính nguy cơ mắc ung thư vú ở cấp độ cá thể, dựa trên các biến nguy cơ kinh điển như tuổi, tuổi có kinh lần đầu, tuổi sinh con đầu lòng, tiền sử sinh thiết vú và tiền sử gia đình mắc ung thư vú ở người thân bậc một.[2] Mô hình Rosner-Colditz được Rosner và Colditz xây dựng từ năm 1996 trên cơ sở dữ liệu Nurses' Health Study theo cách tiếp cận log-incidence, tích hợp thêm nhiều yếu tố sinh sản, nội tiết và nhân trắc như số lần đẻ, chỉ số khối cơ thể, tình trạng mãn kinh và sử dụng hormon.[3,4] Các nghiên cứu tổng quan cho thấy khả năng phân biệt của các mô hình dự báo nguy cơ ung thư vú nhìn chung chỉ ở mức trung bình, với AUC thường dao động khoảng 0,53

Tác giả liên hệ: Bùi Thị Oanh
Bệnh viện K
Email: buioanh1310@gmail.com
Ngày nhận: 05/04/2026
Ngày được chấp nhận: 28/04/2026

- 0,66 ở kiểm định nội bộ và 0,56 - 0,63 ở tập kiểm định độc lập; đồng thời độ nhạy và độ đặc hiệu thay đổi đáng kể giữa các quần thể nghiên cứu.[5] Một số nghiên cứu tại châu Á ghi nhận mô hình Gail có xu hướng đánh giá quá mức nguy cơ ung thư vú nếu chưa được hiệu chỉnh phù hợp với quần thể đích.[10]

Trong bối cảnh đó, việc đánh giá lại hiệu năng của các mô hình dự báo nguy cơ ung thư vú trên dữ liệu thực tế tại Việt Nam là cần thiết. Nghiên cứu này được thực hiện nhằm: (1) đánh giá hiệu năng dự báo của mô hình Gail điều chỉnh và mô hình kết hợp Gail-Rosner-Colditz bằng các thuật toán học máy trên bộ dữ liệu sàng lọc có biến cố hiếm; (2) phân tích các yếu tố liên quan với nguy cơ mắc ung thư vú bằng hồi quy logistic phạt Firth.

II. ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP

1. Đối tượng

Nghiên cứu sử dụng bộ dữ liệu thứ cấp gồm 31.016 phụ nữ tham gia chương trình sàng lọc ung thư vú từ năm 2013 - 2021. Chương trình sàng lọc tiếp nhận phụ nữ từ 40 tuổi trở lên, hoặc phụ nữ từ 35 tuổi trở lên có tiền sử gia đình ung thư vú.

Tiêu chuẩn lựa chọn

- Phụ nữ tham gia chương trình sàng lọc trong thời gian nghiên cứu, có đầy đủ thông tin về biến kết cục và các biến đưa vào mô hình phân tích.

- Các trường hợp nghi ngờ trên sàng lọc phải có thông tin chẩn đoán xác định cuối cùng.

Tiêu chuẩn loại trừ

- Hồ sơ trùng lặp; hồ sơ thiếu thông tin ở biến kết cục hoặc một trong các biến độc lập sử dụng trong mô hình.

Trong quần thể này ghi nhận 19 trường hợp ung thư vú được chẩn đoán xác định.

2. Phương pháp

Thiết kế nghiên cứu: Nghiên cứu phân tích hồi cứu trên cơ sở dữ liệu sàng lọc, kết hợp giữa tiếp cận mô hình dự báo và phân tích dịch tễ học.

Biến số nghiên cứu

Nghiên cứu đánh giá nguy cơ ung thư vú dựa trên hai cấu hình mô hình đã được điều chỉnh theo khả năng sẵn có của dữ liệu: (1) mô hình Gail điều chỉnh; (2) mô hình kết hợp Gail-Rosner-Colditz. Do giới hạn của bộ số liệu, nghiên cứu không thể tái lập đầy đủ cấu trúc nguyên bản của hai mô hình gốc. Các biến được đưa vào phân tích gồm một biến liên tục là tuổi và các biến phân loại/nhị phân đã được mã hóa: tuổi kinh lần đầu (7 – 11; 12 – 13; ≥ 14 tuổi), tiền sử gia đình mắc ung thư (không/có), BMI (< 18,5; 18,5 – 22,9; 23 – 24,9; ≥ 24), số lần đẻ (0; 1; 2; 3), sử dụng thuốc nội tiết (không/có). Tiền sử gia đình mắc ung thư được xác định là có khi đối tượng có mẹ hoặc chị/em gái ruột (người thân bậc một) mắc ung thư vú. Sử dụng thuốc nội tiết được định nghĩa là có tiền sử hoặc đang sử dụng các chế phẩm hormone ngoại sinh, bao gồm liệu pháp hormone thay thế, thuốc hỗ trợ sinh sản hoặc thuốc tránh thai nội tiết.

Mô hình Gail điều chỉnh sử dụng các biến cốt lõi gồm tuổi, tuổi kinh lần đầu và tiền sử gia đình. Mô hình kết hợp Gail-Rosner-Colditz tích hợp thêm số lần đẻ, BMI và tiền sử sử dụng thuốc nội tiết. Nhóm nghiên cứu sử dụng mô hình kết hợp được xây dựng để đánh giá giá trị dự báo gia tăng của nhóm biến này trên bộ dữ liệu sẵn có.

Xử lý và phân tích số liệu

Dữ liệu được phân chia ngẫu nhiên thành tập huấn luyện (60%) và tập kiểm tra độc lập (40%) bằng phương pháp lấy mẫu phân tầng nhằm bảo toàn tỷ lệ hiện mắc ung thư vú ở cả hai tập. Tập kiểm tra được giữ cố định và không tham gia vào quá trình huấn luyện. Do tỷ lệ biến cố trong quần thể nghiên cứu cực kỳ

thấp (0,06%), các mô hình đối mặt với nguy cơ sai số quá khớp (overfitting) rất lớn. Để khống chế nguy cơ này, tập kiểm tra được cô lập hoàn toàn, không tham gia vào bất kỳ bước nào của quá trình tiền xử lý và huấn luyện nhằm ngăn ngừa triệt để hiện tượng rò rỉ dữ liệu (data leakage).

Để khắc phục tình trạng mất cân bằng dữ liệu nghiêm trọng, chúng tôi áp dụng chiến lược tái lấy mẫu kết hợp trên tập huấn luyện: giảm mẫu ngẫu nhiên (random undersampling) nhóm đa số và tăng mẫu nhóm thiểu số bằng kỹ thuật SMOTE (Synthetic Minority Over-sampling Technique - kỹ thuật tăng mẫu tổng hợp cho nhóm thiểu số).[8] Do bản chất của thuật toán SMOTE là nội suy tuyến tính, kỹ thuật này có thể sinh ra các giá trị thập phân đối với các biến phân loại và biến rời rạc. Để khắc phục hạn chế này của phương pháp SMOTE, các giá trị thuộc biến rời rạc đều được áp dụng kỹ thuật làm tròn về số nguyên gần nhất. Bước hiệu chỉnh này đảm bảo tập dữ liệu tổng hợp hoàn toàn hợp lý và tuân thủ chặt chẽ các đặc tính của dữ liệu gốc. Ba thuật toán học máy được triển khai gồm: hồi quy logistic kernel có trọng số lớp (Weighted Kernel Logistic Regression), thuật toán k láng giềng gần nhất cho lớp hiếm (k-Rare-class K-Nearest Neighbors, KRNN) và rừng ngẫu nhiên cân bằng (Balanced Random Forest). Tính ổn định của mô hình được đánh giá bằng bootstrap 100 vòng lặp; tại mỗi vòng, mô hình được huấn luyện lại trên tập huấn luyện và đánh giá trên tập kiểm tra độc lập.

Hiệu năng mô hình được lượng giá bằng ROC-AUC, PR-AUC, chỉ số Brier, độ chính xác dương tính, độ nhạy và chỉ số F1 (thước đo hài hòa giữa độ chính xác dương tính và độ nhạy). Trong bối cảnh dữ liệu mất cân bằng nặng, PR-AUC được xem là chỉ số quan trọng để phản ánh hiệu năng nhận diện nhóm ung thư vú, bên cạnh ROC-AUC.[9]

Song song với các mô hình dự báo, nghiên cứu sử dụng hồi quy logistic phạt Firth để phân tích các yếu tố liên quan với ung thư vú nhằm giảm sai lệch ước lượng trong bối cảnh biến cố hiếm và nguy cơ phân tách dữ liệu.[6,7] Do số biến cố ung thư vú trong bộ dữ liệu rất thấp, việc sử dụng hồi quy logistic thông thường có nguy cơ gặp hiện tượng phân tách dữ liệu và tạo ra ước lượng chệch. Vì vậy, nghiên cứu sử dụng hồi quy logistic phạt Firth nhằm giảm sai lệch ước lượng trong bối cảnh biến cố hiếm và cho phép ước lượng tỷ suất chênh ổn định hơn. Kết quả được biểu diễn bằng tỷ suất chênh hiệu chỉnh (aOR) và khoảng tin cậy 95%. Ngưỡng ý nghĩa thống kê được xác định với $p < 0,05$. Toàn bộ phân tích được thực hiện bằng Python phiên bản 3.9 với các thư viện scikit-learn và imbalanced-learn.

3. Đạo đức nghiên cứu

Nghiên cứu sử dụng cơ sở dữ liệu đã được ẩn danh. Đề cương nghiên cứu được thông qua Hội đồng đạo đức Bệnh viện K với số chấp thuận 3379/QĐ-BVK ngày 16/12/2022.

III. KẾT QUẢ

1. Đặc điểm đối tượng nghiên cứu

Tuổi trung bình của 31.016 phụ nữ trong nghiên cứu là 49,2 ± 7,7 tuổi. Nhóm có tuổi kinh lần đầu ≥ 14 tuổi chiếm 84,7%; 5,2% có tiền sử gia đình mắc ung thư; 61,5% có BMI từ 18,5 đến 22,9; 47,6% đã sinh 2 lần; và 8,7% có tiền sử sử dụng thuốc nội tiết (Bảng 1).

Bảng 1 trình bày phân bố đặc điểm của toàn bộ quần thể và theo tình trạng ung thư vú. Do số trường hợp ung thư vú rất ít (n = 19), các so sánh giữa hai nhóm trong bảng này chỉ mang tính mô tả; mối liên quan được lượng giá bằng hồi quy logistic phạt Firth ở mục 5.

Bảng 1. Đặc điểm đối tượng nghiên cứu (n = 31.016)

Đặc điểm	Toàn bộ	Ung thư	
		Không	Có
	n (%)	n (%)	n (%)
Tuổi, trung bình ± SD	49,2 ± 7,7	49,2 ± 7,7	52,5 ± 9,3
Tuổi kinh lần đầu			
7 – 11	292 (0,9)	292 (0,9)	0 (0,0)
12 – 13	4.449 (14,3)	4.446 (14,3)	3 (15,8)
≥ 14	26.275 (84,7)	26.259 (84,7)	16 (84,2)
Tiền sử gia đình ung thư			
Không	29.389 (94,8)	29.372 (94,8)	17 (89,5)
Có	1.627 (5,2)	1.625 (5,2)	2 (10,5)
BMI			
< 18,5	1.426 (4,6)	1.425 (4,6)	1 (5,3)
18,5 – 22,9	19.088 (61,5)	19.077 (61,5)	11 (57,9)
23 – 24,9	6.964 (22,5)	6.961 (22,5)	3 (15,8)
≥ 24	3.538 (11,4)	3.534 (11,4)	4 (21,1)
Số lần đẻ			
0	1.383 (4,5)	1.376 (4,4)	7 (36,8)
1	2.997 (9,7)	2.992 (9,7)	5 (26,3)
2	14.760 (47,6)	14.754 (47,6)	6 (31,6)
3	11.876 (38,3)	11.875 (38,3)	1 (5,3)
Sử dụng thuốc nội tiết			
Không	28.307 (91,3)	28.290 (91,3)	17 (89,5)
Có	2.709 (8,7)	2.707 (8,7)	2 (10,5)

2. Phân bố đối tượng theo tập dữ liệu và biến cố ung thư vú

Tỷ lệ hiện mắc ung thư vú trong toàn bộ quần thể ở mức rất thấp, chỉ 0,06% (19/31.016). Phân chia phân tầng đã bảo toàn tỷ lệ biến cố tương đồng giữa tập huấn luyện và tập kiểm tra (Bảng 2).

Bảng 2. Phân bố đối tượng nghiên cứu và biến cố ung thư vú

Tập dữ liệu	Tổng số (n)	Không ung thư, n (%)	Ung thư vú, n (%)
Toàn bộ mẫu	31.016	30.997 (99,94)	19 (0,06)
Tập huấn luyện	18.609	18.598 (99,94)	11 (0,06)
Tập kiểm tra	12.407	12.399 (99,94)	8 (0,06)

3. Hiệu năng dự báo của các mô hình học máy

Khi so sánh giữa hai cấu hình mô hình, việc bổ sung các biến của Rosner-Colditz làm tăng chỉ số ROC-AUC của cả ba thuật toán. Weighted Kernel Logistic Regression trong mô hình kết hợp đạt ROC-AUC cao nhất (0,662 ± 0,062). Tuy vậy, PR-AUC của tất cả mô hình vẫn gần bằng 0, phản ánh hạn chế trong nhận diện nhóm ung thư vú trên bộ dữ liệu có biến cố hiếm (Bảng 3).

Bảng 3. Hiệu năng dự báo của các mô hình học máy trên tập kiểm tra

Thuật toán học máy	Mô hình Gail điều chỉnh			Mô hình Gail-Rosner-Colditz		
	ROC-AUC	PR-AUC	Chỉ số Brier	ROC-AUC	PR-AUC	Chỉ số Brier
Weighted Kernel Logistic Regression	0,389 (0,010)	0,000 (0,000)	0,205 (0,002)	0,662 (0,062)	0,001 (0,000)	0,089 (0,002)
k-Rare-class KNN	0,497 (0,006)	0,000 (0,000)	0,002 (0,001)	0,582 (0,051)	0,003 (0,011)	0,024 (0,002)
Balanced Random Forest	0,403 (0,013)	0,000 (0,000)	0,178 (0,002)	0,570 (0,030)	0,003 (0,011)	0,062 (0,002)

4. Hiệu năng phân loại đối với nhóm ung thư vú

Ở nhóm ung thư vú, Weighted Kernel Logistic Regression của mô hình kết hợp đạt độ nhạy cao nhất (0,3750), nhưng độ chính xác dương tính vẫn rất thấp (0,0024). KRNN không ghi nhận trường hợp ung thư vú nào được phân loại đúng ở cả hai cấu hình mô hình (Bảng 4).

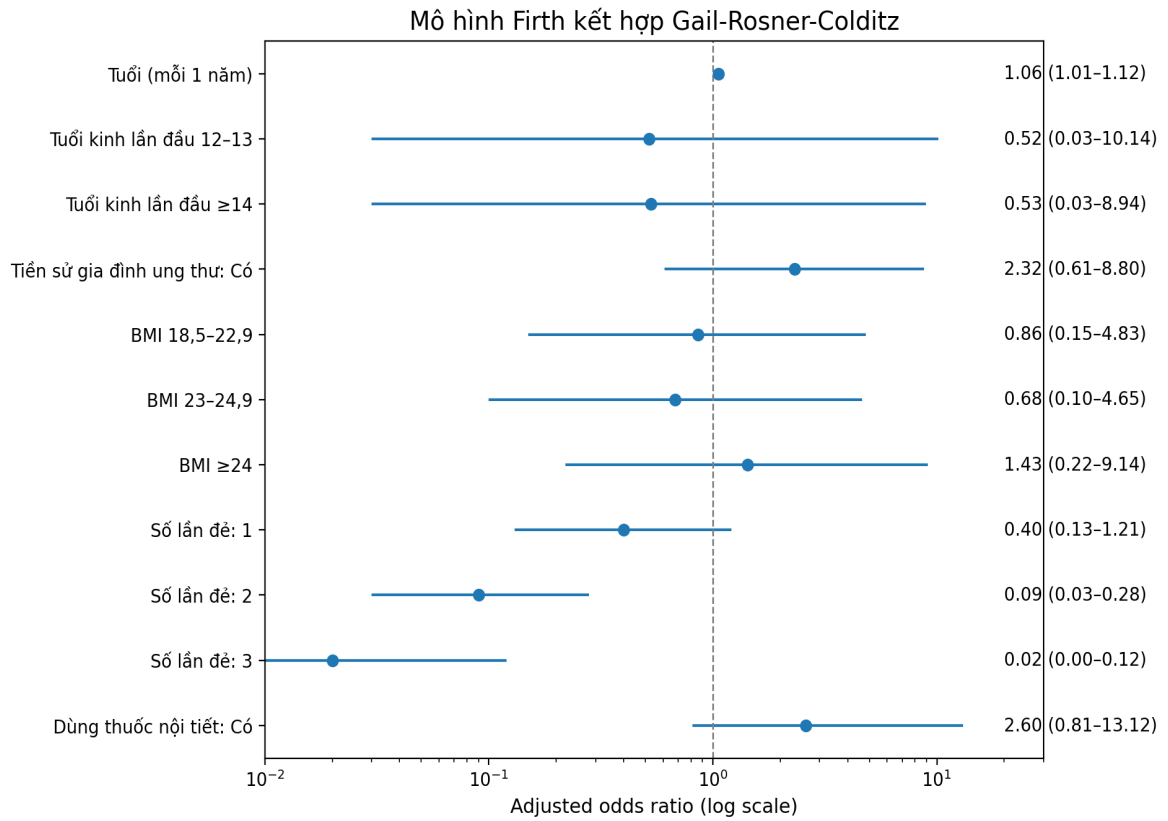
Bảng 4. Hiệu năng phân loại đối với nhóm ung thư vú trên tập kiểm tra

Thuật toán học máy	Mô hình Gail điều chỉnh			Mô hình Gail-Rosner-Colditz		
	Độ chính xác dương tính	Độ nhạy	Chỉ số F1	Độ chính xác dương tính	Độ nhạy	Chỉ số F1
Weighted Kernel Logistic Regression	0,0003	0,1250	0,0006	0,0024	0,3750	0,0047

Thuật toán học máy	Mô hình Gail điều chỉnh			Mô hình Gail-Rosner-Colditz		
	Độ chính xác dương tính	Độ nhạy	Chỉ số F1	Độ chính xác dương tính	Độ nhạy	Chỉ số F1
k-Rare-class KNN	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Balanced Random Forest	0,0006	0,2500	0,0012	0,0025	0,2500	0,0049

5. Hồi quy logistic phạt Firth

Trong mô hình Gail điều chỉnh, tuổi có liên quan với nguy cơ mắc ung thư vú (aOR = 1,05; KTC95%: 1,00 – 1,11), trong khi tuổi kinh lần đầu và tiền sử gia đình mắc ung thư chưa ghi nhận mối liên quan có ý nghĩa thống kê. Trong mô hình kết hợp Gail-Rosner-Colditz, tuổi tiếp tục có liên quan với nguy cơ mắc ung thư vú (aOR = 1,06; KTC95%: 1,01 – 1,12); nhóm sinh 2 con và sinh 3 con có liên quan với giảm nguy cơ mắc so với nhóm chưa sinh con (aOR lần lượt là 0,09 và 0,02) (Biểu đồ 1).



Biểu đồ 1. Biểu đồ rừng của mô hình hồi quy logistic phạt Firth theo cấu hình Gail-Rosner-Colditz

IV. BÀN LUẬN

Kết quả nghiên cứu cho thấy tỷ lệ hiện mắc ung thư vú trong quần thể sàng lọc là 0,06%, phản ánh bộ dữ liệu có biến cố hiếm. Đặc điểm này giải thích vì sao, mặc dù ROC-AUC tăng sau khi bổ sung các biến Rosner-Colditz, PR-AUC và độ chính xác dương tính vẫn ở mức rất thấp. Trong bối cảnh mất cân bằng dữ liệu nặng, chỉ số Brier thấp chủ yếu phản ánh tỷ lệ dự báo âm tính cao, chưa đồng nghĩa với khả năng nhận diện tốt nhóm ung thư vú.[9]

Mô hình kết hợp Gail-Rosner-Colditz cho hiệu năng dự báo cao hơn mô hình Gail điều chỉnh ở cả ba thuật toán, trong đó Weighted Kernel Logistic Regression đạt ROC-AUC cao nhất là 0,662. Giá trị này nằm trong khoảng AUC đã được báo cáo đối với các mô hình dự báo nguy cơ ung thư vú dựa trên bảng hỏi, vốn thường chỉ đạt khả năng phân biệt ở mức trung bình. Trên phụ nữ Hàn Quốc, Min và cộng sự cũng ghi nhận hiệu năng của các mô hình đánh giá nguy cơ có thể thay đổi theo quần thể nghiên cứu, qua đó nhấn mạnh sự cần thiết của việc thẩm định mô hình trên quần thể đích trước khi áp dụng.[10] Kết quả này cũng phản ánh giới hạn của bộ dữ liệu nghiên cứu khi số biến đầu vào còn hạn chế và tổng số trường hợp ung thư vú chỉ là 19.

Phân tích hồi quy logistic phạt Firth cho thấy tuổi có liên quan với nguy cơ mắc ung thư vú; trong khi số lần đẻ có liên quan với giảm nguy cơ mắc trong mô hình kết hợp. Cụ thể, so với nhóm chưa sinh con, nhóm sinh 2 con và nhóm sinh 3 con có aOR lần lượt là 0,09 và 0,02. Xu hướng này phù hợp với cấu trúc lý thuyết của mô hình Rosner-Colditz, trong đó các yếu tố sinh sản tham gia vào ước tính nguy cơ mắc ung thư vú.[3,4]

Trong nghiên cứu này, tuổi kinh lần đầu, tiền sử gia đình, BMI và sử dụng thuốc nội tiết chưa ghi nhận mối liên quan có ý nghĩa thống kê với nguy cơ mắc ung thư vú. Kết quả này nhiều khả năng phản ánh hạn chế về sức mạnh thống kê do số biến cố rất thấp, đồng thời cho thấy bộ dữ liệu hiện có chưa cho phép tái lập đầy đủ các biến nguyên bản của các mô hình gốc. Khoảng tin cậy rộng ở một số biến là biểu hiện của tình trạng dữ liệu thưa.

Điểm mạnh của nghiên cứu là cỡ mẫu quần thể lớn, quy trình tách tập kiểm tra độc lập bằng lấy mẫu phân tầng, có kiểm soát mất cân bằng dữ liệu trên tập huấn luyện và kết hợp đồng thời tiếp cận dự báo với phân tích dịch tễ học. Tuy nhiên, nghiên cứu có một số hạn chế: chỉ có 19 trường hợp ung thư vú nên nguy cơ quá khớp và bất ổn định ước lượng là đáng kể; dữ liệu được thu thập từ một chương trình sàng lọc nên khả năng khái quát hóa còn hạn chế; nhiều biến quan trọng của mô hình Gail và Rosner-Colditz nguyên bản chưa có trong cơ sở dữ liệu; và nghiên cứu chưa thực hiện kiểm định ngoài trên quần thể độc lập.

Từ các kết quả trên, mô hình kết hợp Gail-Rosner-Colditz cho hiệu năng dự báo cao hơn mô hình Gail điều chỉnh trong bộ dữ liệu nghiên cứu, nhưng chưa đủ cơ sở để sử dụng như công cụ dự báo độc lập trong thực hành lâm sàng. Các nghiên cứu tiếp theo cần tăng số biến cố, mở rộng phạm vi thu thập và đánh giá trên quần thể độc lập để xác định giá trị ứng dụng của mô hình.

V. KẾT LUẬN

Trên bộ dữ liệu 31.016 phụ nữ tham gia sàng lọc ung thư vú, mô hình kết hợp Gail-Rosner-Colditz cho hiệu năng dự báo cao hơn mô hình Gail điều chỉnh, trong đó Weighted Kernel Logistic Regression đạt ROC-AUC cao nhất. Tuy nhiên, hiệu năng dự báo đối với nhóm ung thư vú vẫn còn hạn chế do số biến cố rất thấp. Cần tiếp tục thẩm định mô hình trên bộ dữ liệu có số ca bệnh lớn hơn và trên quần thể độc lập trước khi xem xét ứng dụng trong thực hành lâm sàng.

Lời cảm ơn

Nhóm tác giả trân trọng cảm ơn Quỹ Ngày mai tươi sáng đã hỗ trợ trong quá trình sử dụng và xử lý số liệu sàng lọc phục vụ nghiên cứu. Các tác giả cam kết không có xung đột lợi ích

liên quan đến nghiên cứu.

TÀI LIỆU THAM KHẢO

1. Ferlay J, Ervik M, Lam F, et al. *Global Cancer Observatory: Cancer Today*. International Agency for Research on Cancer; 2024.

2. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst*. 1989;81(24):1879-1886. doi:10.1093/jnci/81.24.1879.

3. Rosner B, Colditz GA. Nurses' Health Study: log-incidence mathematical model of breast cancer incidence. *J Natl Cancer Inst*. 1996;88(6):359-364. doi:10.1093/jnci/88.6.359.

4. Rice MS, Tworoger SS, Hankinson SE, et al. Breast cancer risk prediction: an update to the Rosner-Colditz breast cancer incidence model. *Breast Cancer Res Treat*. 2017;166(1):227-240. doi:10.1007/s10549-017-4391-5.

5. Anothaisintawee T, Teerawattananon Y, Wiratkapun C, Kasamesup V, Thakkinstian A. Risk prediction models of breast cancer: a systematic review of model performances. *Breast Cancer Res Treat*. 2012;133(1):1-10. doi:10.1007/s10549-011-1853-z.

6. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika*. 1993;80(1):27-38. doi:10.1093/biomet/80.1.27.

7. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Stat Med*. 2002;21(16):2409-2419. doi:10.1002/sim.1047.

8. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res*. 2002;16:321-357. doi:10.1613/jair.953.

9. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS One*. 2015;10(3). doi:10.1371/journal.pone.0118432.

10. Min JW, Chang MC, Lee HK. Validation of risk assessment models for predicting the incidence of breast cancer in Korean women. *J Breast Cancer*. 2014;17(3):226-235. doi:10.4048/jbc.2014.17.3.226.

Summary

PERFORMANCE EVALUATION OF MODIFIED GAIL AND GAIL-ROSNER-COLDITZ MODELS FOR BREAST CANCER RISK PREDICTION IN VIETNAM

This retrospective study evaluated the predictive performance of a modified Gail model and a combined Gail-Rosner-Colditz model in 31,016 women participating in a breast cancer screening program, including 19 breast cancer cases. Data were split by stratified sampling into training (60%) and testing (40%) sets, and class imbalance was addressed by random undersampling combined with SMOTE. Model performance was assessed using ROC-AUC, PR-AUC, and Brier score; factors associated with breast cancer risk were analyzed using Firth's penalized logistic regression. Adding Rosner-Colditz variables increased ROC-AUC across all three algorithms, and Weighted Kernel Logistic Regression achieved the highest ROC-AUC (0.662 ± 0.062). Age was associated with higher odds of breast cancer, whereas parity of two or three births was associated with lower odds compared with nulliparity. The combined Gail-Rosner-Colditz model showed better predictive performance than the modified Gail model, although overall predictive value remained limited because of the very low number of events.

Keywords: Breast cancer, Gail model, Rosner-Colditz model, risk prediction, machine learning.