Towards Personalized Breast Cancer Risk Management: A Thai Cohort Study on Polygenic Risk Scores

Vorthunju Nakhonsri[1], Manop Pithukpakorn[2,3], Jakris Eu-ahsunthornwattana[4], Chumpol Ngamphiw[1], Rujipat Wasitthankasem[1], Alisa Wilantho[1], Pongsakorn Wangkumhang[1], Manon Boonbangyang[5], Sissades Tongsima[1]*

[1] National Biobank of Thailand (NBT), National Center for Genetic Engineering and Biotechnology (BIOTEC), National Science and Technology Development Agency, Pathum Thani, 12120, Thailand

[2] Division of Medical Genetics, Department of Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand

[3] Siriraj Genomics, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand

[4] Department of Community Medicine, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

[5] Professor Pornchai Matangkasombut Center for Microbial Genomics (CENMIG), Department of Microbiology, Faculty of Science, Mahidol University, Bangkok, Thailand

*Corresponding author: Sissades Tongsima(sissades.ton@biotec.or.th)

Short running title: Population-Specific Validation of Polygenic Risk Scores for Breast Cancer in Thai Women

Abstract

Polygenic Risk Scores (PRS) are now playing an important role in predicting overall risk of breast cancer risk by means of adding contribution factors across independent genetic variants influencing the disease. However, PRS models may work better in some ethnic populations compared to others, thus requiring populaion-specific validation. This study evaluates the performance of 140 previously published PRS models in a Thai population, an underrepresented ethnic group. To rigorously evaluate the performance of 140 breast PRS models, we employed generalized linear models (GLM) combined with a robust evaluation strategy, including Five-fold cross validation and bootstrap analysis in which each model was tested across 1,000 bootstrap iterations to ensure the robustness of our findings and to identify models with consistently strong predictive ability. Among the 140 models evaluated, 38 demonstrated robust predictive ability, identified through > 163 bootstrap iterations (95% CI: 163.88). PGS004688 exhibited the highest performance, achieving an AUROC of 0.5930 (95% CI: 0.5903–0.5957) and a McFadden's pseudo R^2 of 0.0146 (95% CI: 0.0139–0.0153). Women in the 90^{th} percentile of PRS had a 1.83-fold increased risk of breast cancer compared to those within the 30^{th} to 70^{th} percentiles (95% CI: 1.04–3.18). This study highlights the importance of local validation for PRS models derived from diverse populations, demonstrating their potential for personalized breast cancer risk assessment. Model PGS004688, with its robust performance and significant risk stratification, warrants further investigation for clinical implementation in breast cancer screening and prevention strategies. Our findings emphasize the need for adapting and utilizing PRS in diverse populations to provide more accessible public health solutions.

Keywords: Polygenic Risk Scores, Breast Cancer, Thai Population, PRS validation, Genetic Diversity

Introduction

Breast cancer is one of the main causes of death among women all over the world and is a multifactorial disease that depends on genetic and environmental factors [1]. Although some breast cancer cases are associated with strong penetrant mutations in genes such as *BRCA1* and *BRCA2*, most are associated with multiple low penetrant genetic variants [2]. This polygenic nature of breast cancer underscores the need for tools that can accurately assess an individual's cumulative genetic predisposition. Polygenic risk scores (PRS), which aggregate the effects of these numerous common genetic variants, have emerged as a promising tool in this regard. PRS offer a quantitative measure of an individual's genetic predisposition to breast cancer, potentially enabling more targeted screening and prevention strategies [3-4].

While the field of breast cancer PRS research is rapidly expanding, with over 140 models publicly available through repositories like the PGScatalog [5], a critical knowledge gap remains. The majority of these models were developed using data from Western populations, raising concerns about their accuracy and applicability across diverse ethnic groups [6]. Genetic and environmental variations between populations can significantly influence the performance of PRS, highlighting the urgent need for localized validation and adaptation of existing models. Furthermore, there is a lack of research on these models in Asian populations, especially in Southeast Asia. This absence in the development of PRS increases questions on the generalization of the current models to these groups. To fill this gap and facilitate the ability of PRS to accurately estimate breast cancer risk across ethnicities, regional studies, including this one involving a Thai cohort, are important [7-8]. This is crucial to ensure that PRS can effectively assess breast cancer risk in individuals from various backgrounds and ultimately contribute to more equitable and personalized healthcare.

This study aims to evaluate the performance of existing PRS models in a Thai cohort of breast cancer patients, contributing to a more comprehensive understanding of the generalizability and clinical utility of PRS for breast cancer risk assessment in diverse populations.

Materials and Methods

Study Population

This study utilized whole genome sequencing (WGS) data from 184 unrelated Thai women diagnosed with primary breast cancer who were treated at Siriraj Hospital. These data were obtained from previous studies, and the comprehensive case information was recently

published [9]. To focus on the polygenic contribution to breast cancer risk, 38 patients harboring pathogenic or likely pathogenic (P/LP) variants in known breast cancer genes were excluded from the analysis (see Supplementary Table S1). The control group consisted of WGS data from 434 unrelated Thai individuals without cancer (Supplementary Table S2).

Polygenic Risk Score Acquisition and Calculation

A total of 140 harmonized Polygenic Risk Scores (PRS) related to breast cancer (MONDO:0007254) were downloaded from the PGS Catalog on May 27th 2024 [5]. To ensure compatibility with variant call format (VCF) data derived from WGS sequences, these scores were adapted using an in-house pipeline which involved normalizing the effect alleles to the GRCh38 reference genome using the BCFtools plugin `+fixref` [10] and adjusting the weight of the effect alleles aligned with the reference allele by multiplying them by -1. Each PRS was then calculated using the following formula:

$$PGS_i = \sum_{j=1}^{M} \beta_j \times dosage_{ij}$$

where PGS_i represents the polygenic score for the i^{th} individual, β_j is the weight of the alternate allele at the locus j, and $dosage_{ij}$ is the genotype dosage at that locus for the individual i.

Statistical Analysis

To assess the robustness and generalizability of the PRS models, we employed a bootstrap analysis. In each of 1,000 bootstrap iterations, we randomly sampled 128 breast cancer cases and 128 controls to form a training set. Five-fold cross-validation was applied within this training set to identify the best-performing model for each iteration. Model performance was evaluated using McFadden's Pseudo R^2 and the log-likelihood ratio p-value to assess goodness of fit [11]. The Area Under the Receiver Operating Characteristic Curve (AUROC) was calculated to evaluate the discriminatory ability of each model within an independent test set comprising 56 breast cancer patients and 306 controls. Models were ranked based on the frequency of achieving a statistically significant log-likelihood ratio p-value (<0.05) across the 1,000 bootstrap iterations. The final best-performing model was selected based on the average McFadden's Pseudo R^2 and AUROC values across all iterations. All statistical analyses were performed using the R programming environment [12-15].

Language and Computational Tools

This manuscript was refined using the language model ChatGPT for linguistic and structural improvement of the text. [16]

Results

Performance of Polygenic Risk Scores in Predicting Breast Cancer Risk

A comprehensive bootstrap analysis was conducted on 140 PRS models, using 1,000 iterations to evaluate their ability to predict breast cancer status. Results indicated that, on average, each model demonstrated a statistically significant association with breast cancer status in 142.76 out of the 1000 bootstrap iterations (95% confidence interval: 122.57–163.88). A detailed breakdown of the performance of each model across the bootstrap iterations is provided in Supplementary Table S3, and a visual representation of the distribution of significant associations is shown in Figure 1A.

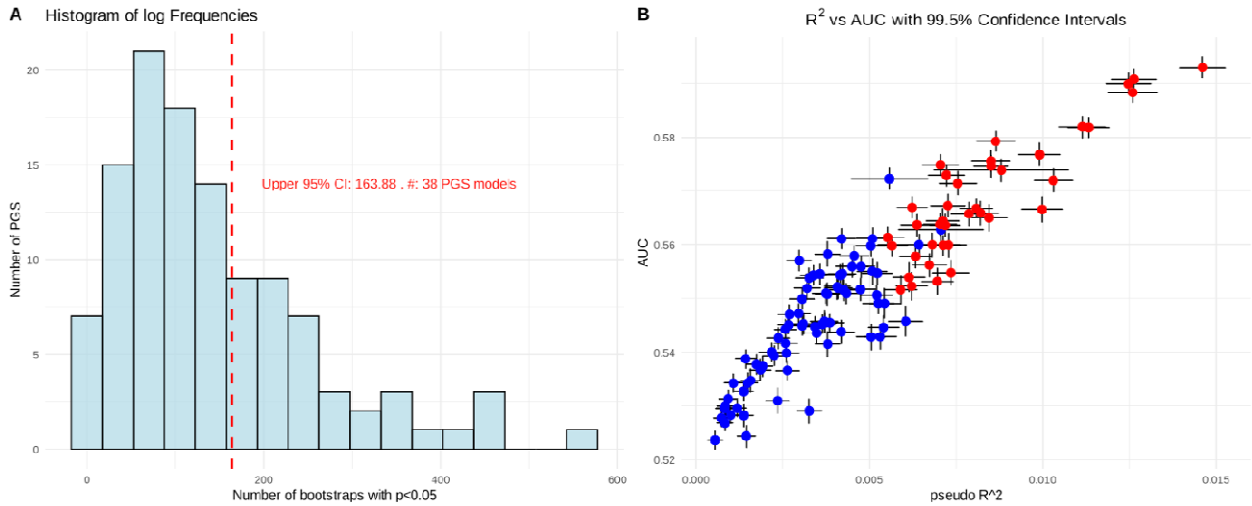

Figure 1: Bootstrap Performance of Polygenic Risk Scores for Breast Cancer Prediction

(A) Distribution of Significant Associations: Histogram displaying the number of bootstrap iterations (out of 1,000) in which each of the 140 PRS models achieved a statistically significant association with breast cancer status (p-value < 0.05). The red dashed line indicates the upper 95% confidence interval (163.88 iterations), highlighting models with frequent significant results. **(B) Predictive Performance and Consistency:** Scatter plot illustrating the relationship between McFadden's Pseudo R^2 and Area Under the Receiver Operating Characteristic Curve (AUROC) for each PRS model. Red dots represent models achieving significance in over 95% of bootstrap iterations, indicating high predictive consistency.

To further evaluate the performance, we plotted McFadden's Pseudo R^2 against AUROC for each model, including 95% confidence intervals (Figure 2A). This analysis identified

PGS004688 as the top-performing model, demonstrating the highest average AUROC (0.5930; 95% CI: 0.5903–0.5957) and Pseudo R^2 (0.0146; 95% CI: 0.0139–0.0153). Figure 2B provides a detailed visualization of the 1,000 bootstrap iterations for PGS004688, with a green square highlighting the mean ± 95% CI of both train and test AUROC values.

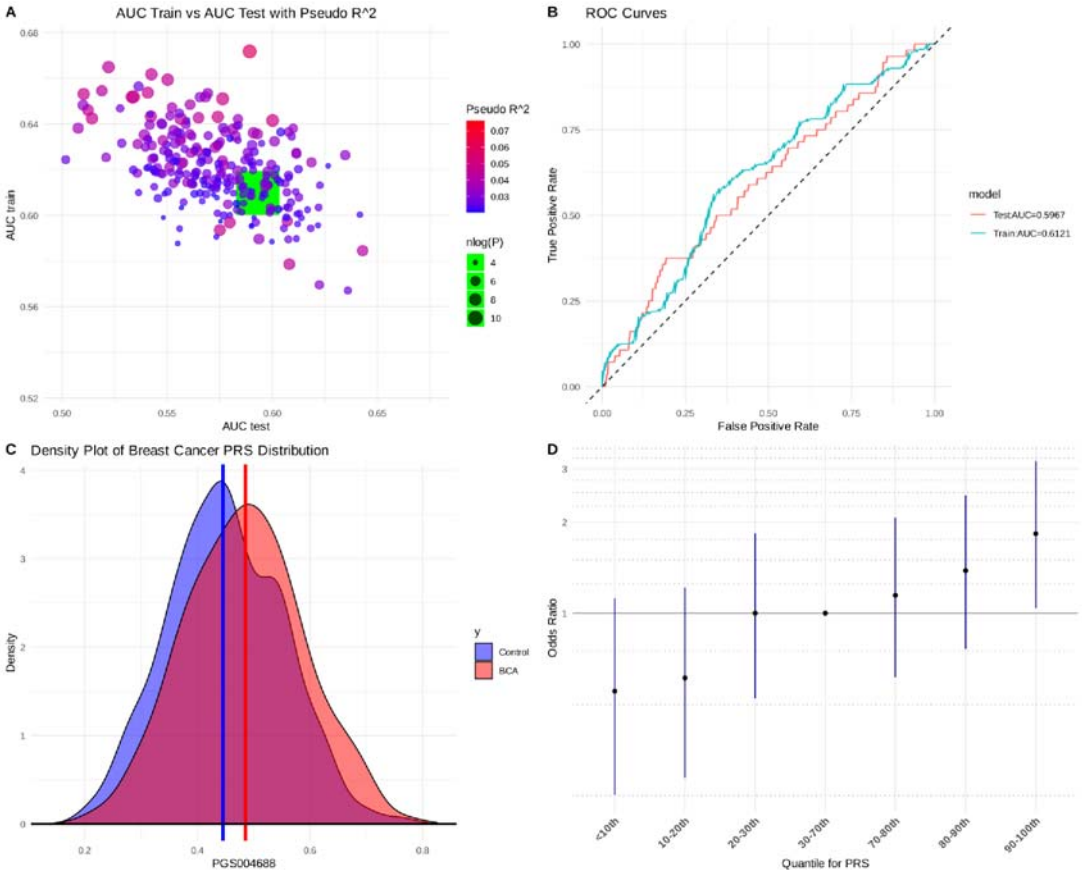

Figure 2: PGS004688: Predictive Accuracy and Consistency

(A) Scatter plot depicting McFadden's Pseudo R^2 versus AUROC for each PRS model. PGS004688 is highlighted, with a green square indicating the mean and 95% confidence interval of AUROC values. **(B)** ROC curves for PGS004688, comparing performance in the training (green) and testing (red) datasets to demonstrate model consistency. **(C)** Density plot illustrating the distribution of standardized PGS004688 scores in breast cancer patients (red) and controls (blue), with median scores indicated. **(D)** Forest plot displaying odds ratios for breast cancer risk at different PGS004688 quantiles. Notably, individuals with scores above the 90th percentile exhibit a significantly elevated risk (odds ratio = 1.83; 95% CI: 1.04–3.18), highlighting the potential clinical utility of PGS004688 for risk stratification.

Discussion

This study underlies the crucial need for population-specific validation of Polygenic Risk Scores (PRS), for accurate breast cancer risk management. Our findings demonstrate that PRS performance can vary significantly across different ethnicities due to variations in genetic diversity and allele frequencies [6]. This discrepancy is particularly evident when comparing European ancestry populations to more genetically diverse populations. While resources like the PGScatalog, containing over 4,000 PRS from over 600 studies, are invaluable, our study highlights the challenges of applying models developed in one population to another.

To address this, we adapted 140 breast cancer-related PRS for use with our Thai cohort. We employed rigorous cross-validation and bootstrap methods to ensure robust model generalization. Notably, we identified PGS004688 as the most effective PRS for predicting breast cancer risk in Thai women. Interestingly, despite being originally developed using GWAS data from a predominantly European cohort [17-18], PGS004688 outperformed models specifically developed for East Asian populations [19-20], This finding underscores the complexity of PRS transferability and the need for population-specific validation. While PGS004688 demonstrated superior performance in our Thai cohort, its effectiveness was lower than its reported performance in European ancestry cohorts (AUROC = 0.665) [18]. This disparity emphasizes the need for continued research and validation of PRS in diverse populations. Further investigation in larger Thai cohorts is crucial to confirm the clinical utility of Ensuring the clinical utility of PGS004688 and ensure its reliability for breast cancer risk assessment in Thailand.

Conclusion

This study highlights the critical need for population-specific validation of Polygenic Risk Score (PRS) for accurate breast cancer risk assessment. Our findings demonstrate that PRS performance can vary significantly across different ethnicities due to variations in genetic diversity and allele frequencies. While resources like the PGScatalog are invaluable, our study reflects the challenges of applying models developed in one population to another. We identified PGS004688 as the most effective PRS for predicting bresat cancer risk in Thai women, outperforming models specifically developed for Eas Asian populations. This finding reveals the complexity of PRS transferability and the need for continued research and validation in diverse populations. Further investigation in larger Thai cohorts is imperative to confirm the clinical utility of PGS004688 and ensure its reliability for breast cancer risk assessment in Thailand.

References

1. World Health Organization. (2024, March 13). Breast cancer. Retrieved from https://www.who.int/news-room/fact-sheets/detail/breast-cancer
2. Gaudet MM, Kirchhoff T, Green T, Vijai J, Korn JM, Guiducci C, et al. Common Genetic Variants and Modification of Penetrance of BRCA2-Associated Breast Cancer. PLoS Genet. 2010 Oct;6(10). doi: 10.1371/journal.pgen.1001183. PMID: 20975944; PMCID: PMC2951372.
3. Lewis, C.M., Vassos, E. Polygenic risk scores: from research tools to clinical instruments. Genome Med 12, 44 (2020). https://doi.org/10.1186/s13073-020-00742-5
4. Roberts E, Howell S, Evans DG. Polygenic risk scores and breast cancer risk prediction. Breast. 2023 Feb;67:71-77. doi: 10.1016/j.breast.2023.01.003. Epub 2023 Jan 10. PMID: 36646003; PMCID: PMC9982311.
5. Lambert, S.A., Gil, L., Jupp, S. et al. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. Nat Genet 53, 420–425 (2021). https://doi.org/10.1038/s41588-021-00783-5
6. Duncan, L., Shen, H., Gelaye, B. et al. Analysis of polygenic risk score usage and performance in diverse human populations. Nat Commun 10, 3328 (2019). https://doi.org/10.1038/s41467-019-11112-0
7. Ho WK, Tai MC, Dennis J, Shu X, Li J, Ho PJ, Millwood IY, Lin K, Jee YH, Lee SH, Mavaddat N, Bolla MK, Wang Q, Michailidou K, Long J, Wijaya EA, Hassan T, Rahmat K, Tan VKM, Tan BKT, Tan SM, Tan EY, Lim SH, Gao YT, Zheng Y, Kang D, Choi JY, Han W, Lee HB, Kubo M, Okada Y, Namba S; BioBank Japan Project; Park SK, Kim SW, Shen CY, Wu PE, Park B, Muir KR, Lophatananon A, Wu AH, Tseng CC, Matsuo K, Ito H, Kwong A, Chan TL, John EM, Kurian AW, Iwasaki M, Yamaji T, Kweon SS, Aronson KJ, Murphy RA, Koh WP, Khor CC, Yuan JM, Dorajoo R, Walters RG, Chen Z, Li L, Lv J, Jung KJ, Kraft P, Pharoah PDB, Dunning AM, Simard J, Shu XO, Yip CH, Taib NAM, Antoniou AC, Zheng W, Hartman M, Easton DF, Teo SH. Polygenic risk scores for prediction of breast cancer risk in Asian populations. Genet Med. 2022 Mar;24(3):586-600. doi: 10.1016/j.gim.2021.11.008. Epub 2021 Dec 15. PMID: 34906514; PMCID: PMC7612481.
8. Ho, WK., Tan, MM., Mavaddat, N. et al. European polygenic risk score for prediction of breast cancer shows similar performance in Asian women. Nat Commun 11, 3833 (2020). https://doi.org/10.1038/s41467-020-17680-w

9. Lertwilaiwittaya P, Roothumnong E, Nakthong P, Dungort P, Meesamarnpong C, Tansa-Nga W, Pongsuktavorn K, Wiboonthanasarn S, Tititumjariya W, Thongnoppakhun W, Chanprasert S, Limwongse C, Pithukpakorn M. Thai patients who fulfilled NCCN criteria for breast/ovarian cancer genetic assessment demonstrated high prevalence of germline mutations in cancer susceptibility genes: implication to Asian population testing. Breast Cancer Res Treat. 2021 Jul;188(1):237-248. doi: 10.1007/s10549-021-06152-4. Epub 2021 Mar 1. PMID: 33649982; PMCID: PMC8233261.
10. Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, Heng Li, Twelve years of SAMtools and BCFtools, GigaScience, Volume 10, Issue 2, February 2021, giab008, https://doi.org/10.1093/gigascience/giab008
11. McFadden D. Conditional logit analysis of qualitative choice behavior. Frontiers in Econometrics. 1973:105–142.
12. R Core Team (2023). _R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
13. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). “Welcome to the tidyverse.” Journal of Open Source Software, *4*(43), 1686. doi:10.21105/joss.01686.
14. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
15. Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez and Markus Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77
16. OpenAI. 2023. "ChatGPT." Accessed June 3, 2023. https://www.openai.com/chatgpt.
17. Zhang H, Ahearn TU, Lecarpentier J, et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. Nature Genetics. 2020 Jun;52(6):572-581. DOI: 10.1038/s41588-020-0609-2. PMID: 32424353; PMCID: PMC7808397.
18. Hu J, Ye Y, Zhou G, Zhao H. Using clinical and genetic risk factors for risk prediction of 8 cancers in the UK Biobank. JNCI Cancer Spectr. 2024 Feb 29;8(2):pkae008. doi: 10.1093/jncics/pkae008. PMID: 38366150; PMCID: PMC10919929.

19. Shieh Y, Hu D, Ma L, Huntsman S, Gard CC, Leung JW, Tice JA, Vachon CM, Cummings SR, Kerlikowske K, Ziv E. Breast cancer risk prediction using a clinical risk model and polygenic risk score. Breast Cancer Res Treat. 2016 Oct;159(3):513-25. doi: 10.1007/s10549-016-3953-2. Epub 2016 Aug 26. PMID: 27565998; PMCID: PMC5033764.
20. Wen W, Shu XO, Guo X, Cai Q, Long J, Bolla MK, Michailidou K, Dennis J, Wang Q, Gao YT, Zheng Y, Dunning AM, García-Closas M, Brennan P, Chen ST, Choi JY, Hartman M, Ito H, Lophatananon A, Matsuo K, Miao H, Muir K, Sangrajrang S, Shen CY, Teo SH, Tseng CC, Wu AH, Yip CH, Simard J, Pharoah PD, Hall P, Kang D, Xiang Y, Easton DF, Zheng W. Prediction of breast cancer risk based on common genetic variants in women of East Asian ancestry. Breast Cancer Res. 2016 Dec 8;18(1):124. doi: 10.1186/s13058-016-0786-1. PMID: 27931260; PMCID: PMC5146840.