RESEARCH ARTICLE

Development and Validation of a Breast Cancer Risk Prediction Model for Thai Women: A Cross-Sectional Study

Thunyarat Anothaisintawee[1,2], Yot Teerawattananon[3], Cholatip Wiratkapun[4], Jiraporn Srinakarin[5], Piyanoot Woodtichartpreecha[6], Siriporn Hirunpat[6], Sansanee Wongwaisayawan[7], Panuwat Lertsithichai[8], Vijj Kasamesup[9], Ammarin Thakkinstian[2]*

Abstract

Background: Breast cancer risk prediction models are widely used in clinical practice. They should be useful in identifying high risk women for screening in limited-resource countries. However, previous models showed poor performance in derived and validated settings. Therefore, we aimed to develop and validate a breast cancer risk prediction model for Thai women. Materials and Methods: This cross-sectional study consisted of derived and validation phases. Data collected at Ramathibodi and other two hospitals were used for deriving and externally validating models, respectively. Multiple logistic regression was applied to construct the model. Calibration and discrimination performances were assessed using the observed/expected ratio and concordance statistic (C-statistic), respectively. A bootstrap with 200 repetitions was applied for internal validation. Results: Age, menopausal status, body mass index, and use of oral contraceptives were significantly associated with breast cancer and were included in the model. Observed/expected ratio and C-statistic were 1.00 (95% CI: 0.82, 1.21) and 0.651 (95% CI: 0.595, 0.707), respectively. Internal validation showed good performance with a bias of 0.010 (95% CI: 0.002, 0.018) and C-statistic of 0.646(95% CI: 0.642, 0.650). The observed/expected ratio and C-statistic from external validation were 0.97 (95% CI: 0.68, 1.35) and 0.609 (95% CI: 0.511, 0.706), respectively. Risk scores were created and was stratified as low (0-0.86), low-intermediate (0.87-1.14), intermediate-high (1.15-1.52), and high-risk (1.53-3.40) groups. Conclusions: A Thai breast cancer risk prediction model was created with good calibration and fair discrimination performance. Risk stratification should aid to prioritize high risk women to receive an organized breast cancer screening program in Thailand and other limited-resource countries.

Keywords: Breast neoplasms - risk prediction model - screening - mammography

Asian Pac J Cancer Prev, **15 (16)**, 6811-6817

Introduction

Breast cancer is the most common female cancer with an incidence of 39/100,000 women (Ferlay et al., 2010). It is also the most common cancer in Thai women with age standardized incidence rate of 25.6/100,000 in the year 2006 (Khuhaprema et al., 2012). Among screening methods, only mammography was efficacious by decreasing mortality rate approximately 20% when compared to non-screening (Nelson et al., 2009; Gotzsche; Nielsen, 2011; Tonelli et al., 2011). Therefore, mammography has been established as an organized screening program in many developed countries (Vainio; Bianchini, 2002), but not for the developing countries due to human resource and infrastructure shortages (Yip et al., 2011).

In Thailand, about 50% of total mammographic machines were dense in the capital city whereas other 46 provinces did not have the mammographic facility (Putthasri et al., 2004). Besides the scarcity of mammographic machine, a number of diagnostic radiologists was also very low that 63 provinces had only one radiologist per province whereas 13 provinces did not have any radiologist (Putthasri et al., 2004). Thus, establishing the organized breast cancer screening program in Thailand is less feasible, but screening in only high risk women may be an alternative for Thailand and other resource limited countries.

Several breast cancer risk prediction models have been developed during the last two decades (Anothaisintawee et al., 2012). These models have been applied to prioritize women for screening, primary chemoprevention

[1]Department of Family Medicine, [2]Section for Clinical Epidemiology and Biostatistics, [4]Department of Radiology, [7]Department of Surgery, [8]Department of Pathology, [9]Department of Community Medicine, Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, [3]Health Intervention and Technology Assessment Program, Ministry of Public Health, Department of Radiology, Faculty of Medicine, [5]Srinagarind Hospital, Khon Kaen University, Khon Kaen, [6]Prince of Songkla University, Songkla, Thailand
**For correspondence: ammarin.tha@mahidol.ac.th*

(Visvanathan et al., 2009), or aid in decision-making between patients and physicians about benefits and harm of breast cancer screening. However, their discriminative performances were poor to fair in either derived (concordance (C) statistic 0.58 to 0.63) or validated settings (C-statistic 0.57 to 0.67) (Anothaisintawee et al., 2012). In addition, most of the models were developed in America, where baseline risks and etiologic factors of breast cancer were different from other countries (Schonfeld et al., 2010). Thus, applying those models to other settings may yield poor performance. We therefore conducted the study aiming to derive and validate the breast cancer risk prediction model in Thai women.

Materials and Methods

Study design and subjects

This cross-sectional study consisted of derived and validation phases. For derived phase, data were consecutively collected during September 2011 to September 2012 at Ramathibodi Hospital, which is a school of medicine hospital located in Bangkok, Thailand. The hospital is also a referral center for complex diseases which serves approximately 5,000 out-patients per day with 1000 in-patient beds. The hospital performs approximately 16,000 mammographic screenings per year.

For the validation phase, data were collected at Srinagarind Hospital in Khon Kaen province, and Songklanagarind Hospital in Songkla province during April 2012 to January 2013. Both hospitals are schools of medicine and referral centers coverage in the North East and South regions, respectively. Women undergoing mammographic screening were eligible if they were aged older than 18 years and agreed to participate. Women with history of invasive breast cancer, ductal carcinoma in situ (DCIS), or other cancers were excluded. The study was approved by Institutional Review Boards of the 3 studied hospitals. All participants provided written informed consents.

Variables and outcome

Well-trained staffs interviewed participants using structured data record forms including demographic data (i.e. age and body mass index (BMI)), risk behavior (i.e. smoking and alcohol intake), family history of breast and ovarian cancers in the first-degree relatives, reproductive data (i.e. age at menarche and at first live birth, breastfeeding, and menopausal status), external hormone usage(i.e. hormonal replacement therapy (HRT), oral contraceptives(OC), and medroxyprogesterone injection), history of breast biopsy, and underlying diseases. External hormone users were defined as follows: current users if they were currently used or used within the last 12 months, past users if they had stopped using longer than 12 months, and never users if they had never or used for less than 1 month. The interested underlying diseases were diabetes mellitus(DM), chronic kidney disease (CKD), and dyslipidemia (DLP). These data were obtained from interviews and subsequently verified with International Classification of Diseases 10 (ICD-10) databases. The primary outcome of interest was combined invasive breast cancer and DCIS confirmed by pathological diagnosis.

Sample size

The prevalence of breast cancer at Ramathibodi Hospital in 2010 was 0.6% (95% confidence interval (CI): 0.5%, 0.7%). Sample size was estimated based on one proportion(Lemeshow, 1990), suggesting 9,845 women were required with type-1 error and CI width of 5% and 0.0015, respectively. According to the recommendation for deriving a prediction model, a valid risk-model required at least 10 subjects with events per one predictor(Guyatt, 2006). If eight significant predictors were expected in the risk model, then 80 breast cancers were required. Given the prevalence of breast cancer of 0.6% and 10% missing data, 15,200 subjects were required in the derive study phase.

Multiple Imputations (MI)

Missing data were imputed using a simulation-based sequential multivariate-regression analysis with chain equations(Rubin and Schenker, 1991; White et al., 2011). Distributions of missing data were explored to determine whether data were missing at random (MAR). Complete data (i.e., diagnosis of breast cancer, age, parity, menopausal status, DM, CKD, DLP, and mammographic results) were used to predict the missing values. Since the frequency of missing data and the largest fraction of missing information (FMI, i.e. uncertainty of the values estimated from MI) were very low in this study (less than 0.05), 10 imputations were efficient to allow for the MI uncertainty (van Buuren et al., 1999; White et al., 2011). Bias from MI was examined using the "midiagplots" command in STATA (Edding and Marchenko, 2012).

Statistical analysis

Derivative phase: data from Ramathibodi hospital were used for deriving the model. A simple logistic regression was applied to assess predictors of breast cancer. Variables with P value less than 0.15 were considered in the multivariate logistic regression. F test with forward elimination was applied to determine the parsimonious model. Goodness of fit of the model was assessed using the Hosmer-Lemeshow test. Calibration coefficients (observed/expected (O/E) ratio) and C-statistics (by a receiver operating characteristic (ROC) curve) were estimated to assess model performances in calibration and discrimination (Harrell et al., 1996).

Coefficients of the significant variables were used to construct a scoring scheme. The risk scores were then assigned to individuals. Total risk scores, a summation of individual scores, were stratified according to the likelihood ratio positive (LR^+) suggested by the ROC curve analysis. Sensitivity, specificity, and LR+ for each score's category were estimated.

Validation phases: the whole data from derived phase was used to internally validate the model using a bootstrap with 200-repetitions (Harrell et al., 1996; Schumacher et al., 1997). For each bootstrap, the derived model was fitted and the probability of breast cancer was estimated. The correlation between observed and predicted values of breast cancer was estimated in the bootstrap data (called Dboot) and derived data (called Doriginal) using

Somer'D coefficient (Harrell et al., 1996). A calibration (called bias) was assessed by subtracting Dboot from Doriginal. The discriminatory performance of our model was evaluated by comparing the original C-statistic with the mean C-statistics from the bootstrap samples.

Data from Srinagarind and Songklanagarind Hospitals were used to externally validate our model. Total scores and the probability of having breast cancer for individuals were calculated based on the derived scoring scheme. The O/E ratio and C-statistic were then estimated.

All analyses were performed using mi estimates in STATA version 12. P value less than 0.05 was considered statistically significant.

Results

Derivative phase

Studied participants: a total of 17,506 women undertook mammographic screening at Ramathibodi Hospital, 15,718 were eligible for the study. Reasons for ineligibility have been described in Figure 1A. Among participating women, the mean age and BMI were respectively 54.98 (± 8.74) years and 23.83 (±3.69) kg/m^2. History of smoking was only 0.66% whereas alcohol intake was 14.96%. About 8.57% and 0.84% of women reported the history of breast and ovarian cancer in their first degree relatives. The prevalence of CKD, DM, and DLP were 1.51%, 8.77%, and 53.03%, respectively. For reproductive history, the mean age at menarche was 14.08 (1.78) years and about 70% were menopauses with mean age at menopause of 48.96 (4.49) years. About 66.5% were parous with mean age at first live birth of 27.64 (4.84) years. A half of them had ever breastfed. The rates of HRT, OC, and medroxyprogesterone injection usage were 17.07%, 27.90%, and 6.29%, respectively. A total of 107 women were diagnosed as breast cancer (invasive breast cancers 91, DCISs=16) with the prevalence of 0.68% (95% CI: 0.56%, 0.82%).

Multiple imputations: thirteen variables (i.e. age at first live birth, BMI, smoking, alcohol intake, family history of breast cancers, breastfeeding, OC usage, history of breast biopsy, family history of ovarian cancer, medroxyprogesterone injection and HRT usages, age at menarche and menopause) contained missing data with the percentage of missing ranging from 0.04% to 7.04%, see Table 1. Exploring distributions of these missing values suggested that missing values were not a sub-set of each other, thus their missing distributions were assumed to be arbitrary-patterns. Therefore, data imputation based on the assumption of MAR could be applied. Distributions of observed and imputed values have been described in Table 1, and suggested that they were very similar for all variables. The diagnostic plot suggested no difference between the missing and observed values, see Figure 1B.

Model selection: distributions of 17 predictors were compared between breast and non-breast cancers. Among them, 12 variables (age, BMI, family history of ovarian cancer, CKD, DM, DLP, duration of breastfeeding, menopausal status, history of breast biopsy, OC, HRT, and medroxyprojesterone injection usages) were considered in the multivariate-logit model. Results of model selection

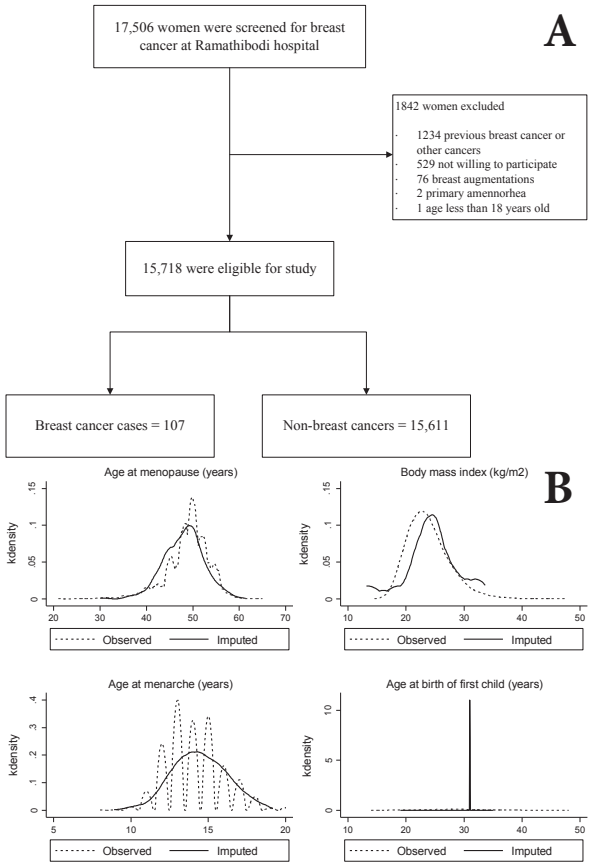

Figure 1. A) Flow Chart of Recruiting Studied Participants; B) Distributions between Missing Values and Observed Values

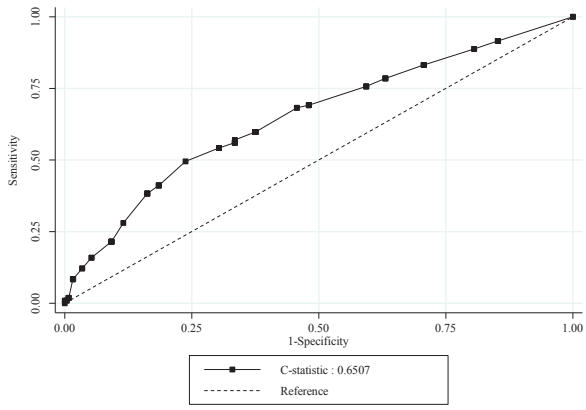

Figure 2. Discriminative Performance of Thai Breast Cancer Risk Prediction Model

indicated only 4 variables (i.e. age, menopausal status, BMI, and OC usage) were significantly associated with breast cancer, they were thus kept in the final model, see Table 2. Women aged older than 60 years were significantly higher risk to have breast cancer than women aged 60 years or younger with the odds ratio (OR) of 1.71 (95% CI: 1.04, 2.81). Current but not past OC users were 4.58 (95% CI: 2.16, 9.71) times significantly higher odds of breast cancer than never users. Obesity, but not overweight, was a significant predictor of breast cancer with OR of 2.02 (95% CI: 1.26, 3.24). Premenopausal women were also significantly higher risk than postmenopausal women with the OR of 1.91 (95% CI: 1.18, 3.08).

Table 1. Describe the Distribution and Frequency of Variables Among Observed and Imputed Datasets

Characteristics	Numbers of missing data (%)	Observe dataset (%)	Imputed dataset (%)	FMI
Demographic data				
BMI, kg/m^2, mean (SD)	12 (0.07)	23.83 (3.69)	23.83 (3.69)	0.007
Family history of breast cancer	21 (0.13)			
No		91.43	91.43	<0.0001
Yes		8.57	8.57	
Family history of ovarian cancer	40 (0.25)			<0.0001
No		99.16	99.16	
Yes		0.84	0.84	
History of breast biopsy	29 (0.18)			<0.0001
No		81.45	81.44	
Yes		18.55	18.56	
History of smoking	16 (0.10)			<0.0001
Never		99.34	99.34	
Ever		0.66	0.66	
History of alcohol drinking	17 (0.11)			<0.0001
Never		85.13	85.12	
Ever		14.87	14.88	
Reproductive history				
Age at menarche, years, mean (SD)	253 (1.60)	14.07 (1.78)	14.08 (1.78)	0.0401
Age at first live birth, years, mean (SD)	4 (0.04)	27.64 (4.84)	28.80 (4.23)	<0.0001
Age at menopause, years, mean (SD)	771 (7.04)	48.96 (4.49)	48.93 (4.49)	0.0368
History of breastfeeding	13 (0.13)			<0.0001
Never		46.09	46.06	
Ever		53.91	53.94	
External hormone use				
History of HRT use	71 (0.65)			0.0535
Never		82.94	82.92	
Ever		17.06	17.08	
History of OC use	28 (0.18)			<0.0001
Never		70.86	70.88	
Ever		29.14	29.12	
History of hormone injection use	44 (0.28)			<0.0001
Never		93.71	93.71	
Ever		6.29	6.29	

$$\ln\left[\frac{P}{1-P}\right] = -5.81 + 0.54 \times Age + 1.52 \times OC_1 + 0.16 \times OC_2 + 0.70 \times BMI_1 + 0.44 \times BMI_2 + 0.64 \times Premen$$

$$\text{probabilit of having breast cancer} = \frac{e^{-5.81+0.54\times Age+1.52\times OC_1+0.16\times OC_2+0.70\times BMI_1+0.44\times BMI_2+0.64\times Premen}}{1+e^{-5.81+0.54\times Age+1.52\times OC_1+0.16\times OC_2+0.70\times BMI_1+0.44\times BMI_2+0.64\times Premen}}$$

Age = 1 if Age > 60 years
Age = 0 if Age ≤ 60 years
OC_1 = 1 if current user of oral contraceptives
OC_2 = 1 if past user of oral contraceptives
OC_1 and OC_2 = 0 if never user of oral contraceptives

BMI_1 = 1 if BMI ≥ 27 kg/m^2
BMI_2 = 1 if BMI = 24 - 26 kg/m^2
BMI_1 and BMI_2 = 0 if BMI ≤ 23 kg/m^2

Premen = 1 if premenopausal women
Premen = 0 if postmenopausal women

Figure 3. Logit Equation of Breast Cancer

Model performance: C-statistic of the final model was 0.651 (95% CI: 0.595, 0.707), see (Figure 2), indicating the model could fairly discriminate breast cancer from non-breast cancer. Hosmer-Lemeshow test indicated the model fitted well with the data (Chi-square test=6.82, P value=0.56) with the O/E ratio of 1.00 (95% CI: 0.82, 1.21).

Scoring scheme: the scoring scheme was constructed based on the estimated coefficients of the 4 variables with the total scores ranged from 0 to 3.4, see Table 2. A probability of breast cancer was estimated using the equation described in Figure 3. The total risk score was stratified into low (0-0.86), low-intermediate (0.87-1.14), intermediate-high (1.15-1.52), and high-risk (1.53-3.40) groups as for suggestion from likelihood ratio positive (LR^+), see Table 3. The LR+ for these 3 later correspond groups were 1.79 (95% CI: 1.50, 2.13), 2.36 (95% CI: 1.85, 3.01), and 5.13 (95% CI: 2.71, 9.70) when compared to low-risk group. The corresponding probabilities of having breast cancer were respectively 0.6%, 0.9%, 1.3%, and 2.5%; and the positive predictive values were 0.54%, 1.21%, 1.59%, and 3.40%, respectively.

Table 2. Factors associated with Breast Cancer and The Scoring Scheme from Multivariate Analysis

Factors	Coefficient	SE	P value	OR (95% CI)	Scoring
Age, year					
>60	0.54	0.25	0.035	1.71 (1.04, 2.81)	0.54
≤60				1	0
OC					
Current user	1.52	0.38	<0.001	4.58 (2.16, 9.71)	1.52
Past user	0.16	0.22	0.473	1.17 (0.76, 1.80)	0.16
Never use				1	0
BMI, kg/m2					
≥27	0.7	0.24	0.003	2.02 (1.26, 3.24)	0.7
24-26	0.44	0.24	0.07	1.55 (0.97, 2.49)	0.44
≤23				1	0
Menopausal status					
Premenopause	0.64	0.24	0.008	1.91 (1.18, 3.08)	0.64
Postmenopause				1	0
Total					0-3.4

Internal validation

The estimated Doriginal and Dboot were 0.301 and 0.292 for the derived and bootstrap models, respectively. The bias was only 0.010 (95% CI: 0.002, 0.018), suggesting good calibration. The C-statistics were 0.651 (0.595, 0.707) and 0.646 (95% CI: 0.642, 0.650) for original and bootstrap models, respectively.

External validation

Data of 4,978 women (n=1,974 and 3,004 for

Table 3. Risk stratification and predictive values of a risk prediction score

Score	No. of breast cancers	No. of non-breast cancers	%Sensitivity (95%CI)	%Specificity (95% CI)	LR+ (95% CI)	%PPV
0-0.86	49	10,873	100	0	1	0.54
0.87-1.14	17	2,202	54.21 (44.33, 63.78)	69.65 (68.92, 70.37)	1.79 (1.50, 2.13)	1.21
1.15-1.52	32	2,280	38.32 (29.23, 48.25)	83.76 (83.16, 84.33)	2.36 (1.85, 3.01)	1.59
1.53-3.40	9	256	8.41 (4.16, 15.79)	98.36 (98.15, 98.55)	5.13 (2.71, 9.70)	3.4

*LR+=likelihood ratio positive; PPV =positive predictive value

Srinagarind and Songklanagarind hospitals) were used for external validation. Among them, 35 women were diagnosed as breast cancer (invasive breast cancer=33, DCIS=2) with the prevalence of 0.70% (95% CI: 0.47%, 0.94%). The derived model worked well in the external dataset with the O/E ratio and the C-statistic of 0.97 (95% CI: 0.68, 1.35) and 0.609 (95% CI: 0.511, 0.706), respectively.

Discussion

The risk prediction model of breast cancer for Thai women was developed using cross-sectional data of women undertaking mammographic screening in Ramathibodi Hospital. The model offers fair discriminative performance with C-statistic of 0.651 and provides good calibration performance with O/E ratio of 1.00. The internal validation indicates good calibration performance with the minimal bias of 0.010 and the C-statistic of 0.646. The model retains similar performance in external validation using cross-sectional data of women in two other university hospitals situated in the different parts of country (with C-statistic of 0.609).

The newly developed model includes age, the current status of using OC, obesity, and premenopausal status. Almost of these variables were not included in the previous models constructed in the U.S.(Gail et al., 1989; Rosner and Colditz, 1996) and most of variables in the previous model were not included in our model for Thai women. The dissimilarity of factors between our and previous models may be explained by the following reasons. Firstly, all previous models were developed based on data from mainly Caucasian populations, while our model was constructed in Asian women. The difference in natural history of breast cancer between these two populations is widely recognized. For example, breast cancer commonly occurs in premenopausal Asians whereas it is more common in postmenopausal Caucasians (Han et al., 2004; Son et al., 2006; Yip, 2009; Keramatinia et al., 2014; Wu et al., 2014). This corresponded to our finding in which premenopausal women were approximately 90 percent higher odds of having breast cancer than post-menopausal women. Secondly, the distribution of disease subtypes according to hormone receptor (i.e. estrogen receptor(ER) and progesterone receptor (PR)) is diverse between Asian and Caucasian women. The ER+ tumor in Asian women is not as common as their Caucasian counterparts (Wiechmann et al., 2009; Telli et al., 2011; Chuthapisith et al., 2012). Previous evidences showed that ER+ tumor was associated with reproductive history (i.e., age at menarche, parity, and breastfeeding) while ER- tumor was not(Althuis et al., 2004; Tsakountakis et al., 2005). This might be a reason why reproductive variables (i.e., age at menarche, parity, and breastfeeding) were not included in our model.

Although our model was well calibrated (O/E ratio=1.00), its discriminative performance was modest (C-statistic=0.651) but still better than previous models (Anothaisintawee et al., 2012). This may be explained by the fact that all variables included in the model were cross-sectionally measured. It would enhance the model performance if time-varying variables for age, use of OC or HRT and durations, and BMI were considered. Nevertheless, these variables are not only difficult to collect due to limitation of recall memory, but also are not practical for a screening tool.

It may be possible to improve the discriminative performance by including biomarker risk factors such as microRNAs(Heneghan et al., 2010) or BRCA genes. However, including biomarker is not our aim that is to develop a simple model for screening women for further mammography in resource limited settings like Thailand. In addition, our model is superior to the previous models developed outside Thailand in terms of the C-statistics (i.e. Gail model =0.58, Rosner&Colditz model=0.63) (Anothaisintawee et al., 2012).

Family history of breast cancer in first degree relative was an established risk factor (Gokdemir-Yazar et al., 2014) which was included in the Gail model. However, this factor was not significantly identified by our model, which might be due to low proportion of family history of breast cancer in our data. A family history of breast cancer in first degree relative should be one of the definite criteria for undertaking routine mammography screening, although it was not significant in our setting.

Our risk prediction model should be useful in the countries with limited resources in prioritizing women for receiving limited mammography services and subsequently reduce the burden of breast cancer in the countries. Although, the LR+ of high-risk category (score: 1.53-3.40) was only 5.13, it provides the important change for post-test probability of having disease as for Users' Guide of Evidence-Based Medicine (Letelier et al., 2008). Suppose that the prevalence of breast cancer in Thai women is 0.68%, then the post-test probability of having breast cancer in high-risk woman is increased to 3.43%. Therefore, women in the high-risk group should have priority to receive a mammographic screening.

This approach aligns with a recent concern on the potential benefits and harms of universal mammography screening in western countries, especially the issue of over-diagnosis which is defined as the "diagnosis of a condition that would never cause symptoms or death during a patient's lifetime" (Kirwan, 2013). The over-diagnosis currently causes unnecessary invasive investigations and treatments in those settings. This

prompts attention on the use of mammography screening only for those with high risk (Beckmann et al., 2014). Also, it is estimated that if this screening tool is applied for all Thai women aged 40-59 years, it can reduce the need for mammography from 7,601,145 per annum (in the case of universal mammography screening) to 126,939 or 264,520 per annum (if our model is used as an initial screening and mammography is provided for those with high-risk or intermediate-high risk groups). As a result, our model has good potential to minimize inequity in assessing breast cancer screening in the Thai health care setting where only the better off are currently undertaking the service, but the poor and high-risk population are left out. However, it is necessary that strong empirical evidence such as those derived from a cluster randomized trial of this screening tool should be demonstrated before the tool is widely introduced as a nation-wide program.

Our study has some strength. Our model was both internally and externally validated according to a recommendation for constructing a clinical prediction score (Altman and Royston, 2000). The co-variables considered in the model were collected based on suggestion from systematic review and meta-analysis of risk factors of breast cancer (Liao et al., 2011; Anothaisintawee et al., 2013; Gao et al., 2013; Sangrajrang et al., 2013); thus missing important variable should be less likely. The sample size of our study was large considering in the setting which did not have a well-constructed data registry, and we had consecutively collected the data by well-trained interviewers using standardized data record forms. Although missing data in our study was minimal, we had applied multiple imputations with chain equation to impute the missing data. The four variables included in the model are all non-invasive, low cost, easy to measure, and available in routine practice. However, there are some limitations. Firstly, this study is cross-sectional, meaning that significant associations between variables and breast cancer cannot be claimed as causal relationships. Secondly, the data from women screened for breast cancer at Ramathibodi Hospital, which is a tertiary hospital were used for deriving the model. Therefore, these data might not be representative for the general Thai women due to the referral bias. Finally, an external validation was performed only in the tertiary hospitals in Southern and the North-eastern regions, which did not cover other levels of health facilities and regions of Thailand.

Our breast cancer risk prediction model has good calibration and fair discriminative performance. Women classified as high or intermediate-high risk should be prioritized to receive the organized breast cancer screening.

Acknowledgements

This study was supported by the Health Intervention and Technology Assessment Program, the Thai Health Promotion Foundation, the Health Systems Research Institute, the Bureau of Policy and Strategy of the Ministry of Public Health, and Thai Health-Global Link Initiative Project. This study was a part of Thunyarat Anothaisintawee's dissertation for Ph. D in Clinical Epidemiology, Faculty of Medicine, Ramathibodi Hospital, Mahidol University. The authors declare that there is no conflict of interests regarding the publication of this paper.

References

Althuis MD, Fergenbaum JH, Garcia-Closas M, et al (2004). Etiology of hormone receptor–defined breast cancer: a systematic review of the literature. *Cancer Epidemiol Biomarkers Prev*, **13**, 1558-68.

Altman DG, Royston P (2000). What do we mean by validating a prognostic model? *Stat Med*, **19**, 453-73.

Anothaisintawee T, Teerawattananon Y, Wiratkapun C, et al (2012). Risk prediction models of breast cancer: a systematic review of model performances. *Breast Cancer Res Treat*, **133**, 1-10.

Anothaisintawee T, Wiratkapun C, Lerdsitthichai P, et al (2013). Risk factors of breast cancer: a systematic review and meta-analysis. *Asia Pac J Public Health*, **25**, 368-87.

Beckmann KR, Roder DM, Hiller JE, et al (2014). Influence of mammographic screening on breast cancer incidence trends in South Australia. *Asian Pac J Cancer Prev*, **15**, 3105-12.

Chuthapisith S, Permsapaya W, Warnnissorn M, et al (2012). Breast cancer subtypes identified by the ER, PR and HER-2 status in Thai women. *Asian Pac J Cancer Prev*, **13**, 459-62.

Edding W, Marchenko Y (2012). Diagnostics for multiple imputation in Stata. *The Stata Journal*, **12**, 353-67.

Ferlay J, Shin H-R, Bray F, et al (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer*, **127**, 2893-917.

Gail M, Brinton L, Byar D, et al (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst*, **81**, 1879-86.

Gao Y, Huang YB, Liu XO, et al (2013). Tea consumption, alcohol drinking and physical activity associations with breast cancer risk among Chinese females: a systematic review and meta-analysis. *Asian Pac J Cancer Prev*, **14**, 7543-50.

Gokdemir-Yazar O, Yaprak S, Colak M, et al (2014). Family history attributes and risk factors for breast cancer in Turkey. *Asian Pac J Cancer Prev*, **15**, 2841-6.

Gotzsche PC, Nielsen M (2011). Screening for breast cancer with mammography. *Cochrane Database Syst Rev*, 1877.

Guyatt GH (2006). Determining prognosis and creating clinical preiction rules. In 'Clinical Epidemiology: How to do clinical practice research', Eds Lippincott Williams & Wilkins, the United States, 323-55

Han W, Kim SW, Park IA, et al (2004). Young age: an independent risk factor for disease-free survival in women with operable breast cancer. *BMC Cancer*, **4**, 82.

Harrell FE, Jr., Lee KL, Mark DB (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*, **15**, 361-87.

Heneghan HM, Miller N, Lowery AJ, et al (2010). Circulating microRNAs as novel minimally invasive biomarkers for breast cancer. *Ann Surg*, **251**, 499-505

Keramatinia A, Mousavi-Jarrahi SH, Hiteh M, et al (2014). Trends in incidence of breast cancer among women under 40 in Asia. *Asian Pac J Cancer Prev*, **15**, 1387-90.

Kirwan CC (2013). Breast cancer screening: what does the future hold? *BMJ*, **346**, 87.

Lemeshow S (1990). The two-sample problem. In 'Adequacy sample size in health study', Eds J. Wiley, West Sussex, England, 9-11

Letelier LM, Rada G, Capurro D, et al (2008). Examples of Likelihood Ratios. In 'Users' Guides to the Medical Literature: A Manual For Evidence-Based Clinical Practice', Eds McGraw-Hill Pofessional, New York, NY, 449-80

Liao S, Li J, Wei W, et al (2011). Association between diabetes mellitus and breast cancer risk: a meta-analysis of the literature. *Asian Pac J Cancer Prev*, **12**, 1061-5.

Nelson HD, Tyne K, Naik A, et al (2009). Screening for breast cancer: an update for the U.S. Preventive Services Task Force. *Ann Intern Med*, **151**, 727-37, 237-42.

Putthasri W, Tangcharoensathien V, Mugem S, et al (2004). Geographical distribution and utilization of mammography in Thailand. *Regional Health Forum*, **8**, 84-91.

Rosner B, Colditz GA (1996). Nurses' health study: log-incidence mathematical model of breast cancer incidence. *J Natl Cancer Inst*, **88**, 359-64.

Rubin DB, Schenker N (1991). Multiple imputation in health-care databases: an overview and some applications. *Stat Med*, **10**, 585-98.

Sangrajrang S, Chaiwerawattana A, Ploysawang P, et al (2013). Obesity, diet and physical inactivity and risk of breast cancer in Thai women. *Asian Pac J Cancer Prev*, **14**, 7023-7.

Schonfeld SJ, Pee D, Greenlee RT, et al (2010). Effect of changing breast cancer incidence rates on the calibration of the Gail model. *J Clin Oncol*, **28**, 2411-7.

Schumacher M, Hollander N, Sauerbrei W (1997). Resampling and cross-validation techniques: a tool to reduce bias caused by model building? *Stat Med*, **16**, 2813-27.

Son BH, Kwak BS, Kim JK, et al (2006). Changing patterns in the clinical characteristics of Korean patients with breast cancer during the last 15 years. *Arch Surg*, **141**, 155-60.

Telli ML, Chang ET, Kurian AW, et al (2011). Asian ethnicity and breast cancer subtypes: a study from the California Cancer Registry. *Breast Cancer Res Treat*, **127**, 471-8.

The benefits and harms of breast cancer screening: an independent review. *The Lancet*, **380**, 1778-86.

Tonelli M, Connor Gorber S, Joffres M, et al (2011). Recommendations on screening for breast cancer in average-risk women aged 40-74 years. *CMAJ*, **183**, 1991-2001.

Tsakountakis N, Sanidas E, Stathopoulos E, et al (2005). Correlation of breast cancer risk factors with HER-2/neu protein overexpression according to menopausal and estrogen receptor status. *BMC Womens Health*, **5**, 1.

Vainio H, Bianchini F (eds.) (2002). IARC Handbooks of cancer prevention, lyon, france: IARCPress.

van Buuren S, Boshuizen HC, Knook DL (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med*, **18**, 681-94.

Visvanathan K, Chlebowski RT, Hurley P, et al (2009). American society of clinical oncology clinical practice guideline update on the use of pharmacologic interventions including tamoxifen, raloxifene, and aromatase inhibition for breast cancer risk reduction. *J Clin Oncol*, **27**, 3235-58.

White IR, Royston P, Wood AM (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med*, **30**, 377-99.

Wiechmann L, Sampson M, Stempel M, et al (2009). Presenting features of breast cancer differ by molecular subtype. *Ann Surg Oncol*, **16**, 2705-10.

Wu LZ, Han RQ, Zhou JY, et al (2014). Incidence and mortality of female breast cancer in Jiangsu, China. *Asian Pac J Cancer Prev*, **15**, 2727-32.

Yip CH, Cazap E, Anderson BO, et al (2011). Breast cancer management in middle-resource countries (MRCs): Consensus statement from the Breast Health Global Initiative. *Breast,* 12-9.

Yip CH (2009). Breast cancer in Asia. *Methods Mol Biol*, **471**, 51-64.