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

Thunyarat Anothaisintawee1,2, Yot Teerawattananon3, Cholatif Wiratkapun4, Jiraporn Srinakarin5, Piyanoot Woodtichartpreech6, Siriporn Hirunpat6, Sansanee Wongwaisayawan7, Panuwat Lertsithichai8, Vijj Kasamesup9, Ammarin Thakkinstian2*

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), 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

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...
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 “mdiagplots” 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/m². 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 menopausal 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. For other cancers, the prevalence 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, menopausal status, 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 medroxyprogesterone injection usages) were considered in the multivariate-logit model. Results of model selection 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 4.78 (95% CI: 2.68, 8.53). Premenopausal women were also significantly higher risk than postmenopausal women with the OR of 1.91 (95% CI: 1.18, 3.08).
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.

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**

<table>
<thead>
<tr>
<th>Score</th>
<th>No. of breast cancers</th>
<th>No. of non-breast cancers</th>
<th>%Sensitivity (95%CI)</th>
<th>%Specificity (95% CI)</th>
<th>LR+ (95% CI)</th>
<th>9PPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0-0.86</td>
<td>49</td>
<td>10,873</td>
<td>100</td>
<td>0</td>
<td>1</td>
<td>0.54</td>
</tr>
<tr>
<td>0.87-1.14</td>
<td>17</td>
<td>2,202</td>
<td>54.2 (44.33, 63.78)</td>
<td>69.65 (68.92, 70.37)</td>
<td>1.79 (1.50, 2.13)</td>
<td>1.21</td>
</tr>
<tr>
<td>1.15-1.52</td>
<td>32</td>
<td>2,280</td>
<td>38.32 (29.23, 48.25)</td>
<td>83.76 (83.16, 84.33)</td>
<td>2.36 (1.85, 3.01)</td>
<td>1.59</td>
</tr>
<tr>
<td>1.53-3.40</td>
<td>9</td>
<td>256</td>
<td>8.41 (4.16, 15.79)</td>
<td>98.36 (98.15, 98.55)</td>
<td>5.13 (2.71, 9.70)</td>
<td>3.4</td>
</tr>
</tbody>
</table>

*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 postmenopausal 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 (Anothaiasantawee 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 investigation and treatments in those settings. 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

DOI:http://dx.doi.org/10.7314/APJCP.2014.15.16.6811
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


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