RESEARCH ARTICLE

Editorial Process: Submission:08/01/2021 Acceptance:12/12/2021

Data Driven for Early Breast Cancer Staging using Integrated Mammography and Biopsy

Tongjai Yampaka*, Duangjai Noolek

Abstract

Objective: Breast cancer patients who have a rapid diagnosis have been better prognosis than late diagnosis. The popular screening is mammogram or ultrasound. In recent years, researchers try to develop data driven models to predict early cancer staging from the first screening. However, data elements are not complete such as lymph node status. Therefore, the Integrated dataset approach will be challenging. **Methods:** Because the data elements are not collected from the same source, joining between mammography and biopsy data were performed using latent variables that determine by tumor severity. The datasets consist of 445 mammography reports and 183 pathological reports. The latent variables of the mammogram dataset were determined by the severity of mass, while latent variables of the pathological dataset were determined by TNM Staging. The latent variables were used to join between two datasets. Then, the prediction models were built using the machine learning technique. The modeling is divided into three steps; staging prediction, lymph node prediction, and prognosis. **Results:** Integrated dataset from mammography and biopsy extend more factors and built the models to predict breast cancer staging in the mammography process. The staging prediction is 100% accuracy. The lymph node prediction are 72.47% accuracy, 73.94% specificity, and 72.5% sensitivity. An area under ROC curve is 0.74. The prognosis model prediction are 72.72% accuracy, 80% specificity, and 77% sensitivity. An area under ROC curve is 0.87. There are also built the rule for early staging, diagnosis, and prognosis. **Conclusion:** This study aims to build the models for early staging, diagnosis, and prognosis using the less aggressive method. The advantages are (1) predict staging from the first screening (2) estimate the lymph node metastases for planning to ALND or SLNB (3) evaluate overall survival time. These advantages help the physician planning the best treatment for cancer patients.

Keywords: Breast cancer screening- breast cancer prognosis- breast cancer staging

Asian Pac J Cancer Prev, **22 (12)**, 4069-4074

Introduction

Breast cancer is the leading of cancer in Thailand (NCI Statistic, 2019). The leading of new cancer patients in Stage II and Stage III is over 50%. Cancer staging is a significant factor to predict the outcomes and prognosis. The stage 0 describing non-invasive cancers that remain within their original location and stage IV describing invasive cancers that have spread outside the breast to other parts of the body. Consequently, rapidly determining cancer staging has become a necessity because it helps the physician to determine this during surgery to remove cancer and look at one or more of the underarm lymph nodes. In recent years, many researchers developed prediction models for cancer screening to classify benign or malignant breast lesions. Data driven using machine learning techniques has been proposed in many medical researches (Gouda et al., 2012; Ahmad et al., 2015). In many years, a variety of techniques such as Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Decision Trees (DTs) have been widely applied in cancer research (Konstantina, 2015). These studies showed more improving accuracy in breast cancer diagnosis. The application of Bayesian (Ogunsakin and Siaka, 2017) was produced precise estimates for modeling malignant breast cancer. Their results suggest that age and women with at least high school education have a higher risk of being diagnosed with malignant breast tumors than benign breast tumors. Clinical decision support system proposed in Linqi et al., (2016) is capable of helping physician make diagnosis decision using contextual learning of demographics and medical history data to improve diagnosis accuracy. In addition, the prediction model performance is important. The main problem of machine learning is insufficient samples to learn all possible patterns. Especially, the medical samples are hardly accessible for researchers. Other problems, data elements are not complete from screening to ending outcome. To address these problems, Integrated dataset approach will be challenging.

*Department of Computer Science, Faculty of Business Administration and Information Technology, Rajamangala University of Technology Tawan-Ok, Thailand. *For Correspondence: tongjai_ya@rmutto.ac.th*

Relational join method is generally combined using two common key. Furthermore, the common keys are not visible for joining between two datasets such as mammography images and pathological reports. The semantic join technique (Yeye et al., 2015) was proposed to join the relationship from semantic instead of common keys. Inspired by this technique, the semantic of severity breast lesion is introduced for integrated dataset.

Data driven for early breast cancer staging using integrated mammography and biopsy aims to build the models for early staging, diagnosis, and prognosis using the less aggressive method using latent variables. The mammogram dataset shows only tumor characteristic, while clinical pathological shows proteomic tumor analysis, lymph node status, ER status, PR Status, HER2 Status and overall survival time (OS). Consequently, rapidly determining cancer staging has become a necessity because it helps the physician to determine this during surgery to remove cancer and look at one or more of the underarm lymph nodes. The prediction process divides into three models; lymph node prediction, staging prediction, and prognosis prediction (to estimate overall survival time). The remaining of this paper is assigned as follows. Section 2 presents the materials and methodology. Section 3 shows the experiments and the results, then conclusion followings in section 4.

Materials and Methods

Materials

Mammogram dataset was provided on UCI Machine Learning Repository. It consists of 515 (53.6 %) benign and 446 (46.4 %) malignant including tumor characteristic features. BI-RADS assessment is standard breast imaging report. It is categorized in 1 to 5. There are negative, benign, probably benign, suspicious, and highly suggestive for malignancy respectively. Mass shape is categorized in 1 to 4. There are round, oval, lobular and irregular respectively. If the mass like a purely round or oval shape, it is likely benign mass, while lobular or irregular are suspicious for breast cancer. Mass density is categorized in 1 to 4. There are fat-containing, low, iso, and high respectively. The density relates with the expected attenuation of an equal volume of fibroglandular tissue. High density is associated with malignancy. Mass margin is categorized in 1 to 5. There are Well-defined, Obscured, Microlubulated, ill-defined, and speculated respectively. The margin refers to the distance between a tumor and the edge of the surrounding tissue. The tumor with surrounding tissue is rolled in a special ink so that the outer edges, or margins, are clearly visible under a microscope. A pathologist checks the tissue under a microscope to see if the margins are free of cancer cells. Severity is the outcome (class label). There are benign (0) or malignant (1).

Breast Cancer Proteomes dataset was generated from Clinical Proteomic Tumor Analysis Consortium (NCI/NIH). The samples divide breast cancer patients into separate sub-classes and staging including proteomic tumor analysis, lymph node status, ER status, PR Status, HER2 Status and overall survival time (OS).

Methods

Latent variable construction: Latent variables are not directly observed but they are rather inferred from other variables (Ralph, 2014). This section explains that how to construct each latent variable from dataset.

Mammogram latent variable construction: The latent variables of mammogram dataset were determined by severity of mass according to mass shape, density, margin, and BIRADS. All feature values were defined by the order of suspicious mass. Low feature values are possible benign, while high feature values are possible malignant. The latent variables were calculated severity scores by the summation of feature values, then the summation values discrete in three intervals as follow:

$$Serverity\ score = \sum_{i=1}^{n} x_i$$

where n is the number of features, and X_i is the featurei Severity discretization

$$Severity = \begin{cases} 1\ if\ severity\ scores\ \leq\ 7 \\ 2\ if\ severity\ scores\ \leq\ 11 \\ 3\ if\ severity\ scores\ \geq\ 15 \end{cases}$$

Clinical pathological latent variable construction

While tumor size and lymph node status were appeared in this dataset, TNM cancer staging was defined and used to represent the latent variables. TNM scores are tumor size, lymph node appearing, and metastases (Jingming et al., 2017) that follow by The Union for International Cancer Control (UICC) based on anatomic clinical pathological information. TNM staging was defined in:

$$Stage = \begin{cases} 1\ if\ tumor\ size\ <\ 2\ and\ lymph\ node\ 0 \\ 2\ if\ tumor\ size\ \leq\ 5\ and\ lymph\ node\ \leq\ 3 \\ 3\ if\ tumor\ size\ >\ 5\ and\ lymph\ node\ \leq\ 9 \\ 4\ if\ tumor\ size\ >\ 5\ or\ lymph\ node\ >\ 10 \end{cases}$$

The linear regression was used to test these latent variables. If the R-square is close to 1, it demonstrated these latent variables are consistent.

Dataset fusion using Join Method

The previous section, latent variables were constructed and prepared to join. Cartesian product operation was used to merge the record between the mammogram dataset and the clinical pathological dataset through latent variables. A formal definition of the Cartesian product from set theoretical principles follows from a definition of ordered pair. The most common definition of ordered pairs is (x, y) = {{x}, {x, y}}. Under this definition, (x, y) is an element of P (P (X U Y)); P is power set.

$$X \times_\theta Y = \{x_1, x_2, \dots, x_n, y_1, y_2, \dots, y\};$$

Where:

$$\{x_1, x_2, \dots, x_n \in X\}\ and\ \{y_1, y_2, \dots, y_n \in Y\}$$

θ is the equality latent variable X and Y.

Prediction Modelling using Machine Learning

Rule induction technique was used to modelling because the rule sets are easy to understand. The rule induction operator is similar to the propositional rule learner (Rapid Miner User Manual, 2014). Starting with the less prevalent classes, the algorithm iteratively grows and prunes rules until there are no positive examples left or the error rate is greater than 50%. In the growing phase, for each rule greedily conditions are added to the rule until it is perfect (i.e. 100% accurate). The procedure tries every possible value of each attribute and selects the condition with highest information gain.

k-fold cross-validation: The random sampling was used to avoid the training bias. The integrated dataset was divided into 10-fold cross-validation subsets. The models performances were evaluated the classification accuracy, sensitivity, and specificity. True positive (TP) is correctly identified when the disease is occurring, while true negative (TN) is correctly identified when the disease is not occurring. In other hand, False positive (FP) is incorrectly identified when the disease is occurring, while false negative (FN) is incorrectly identified when the disease is not occurring. The measure metrics as:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

$$Sensitivity = \frac{TP}{(TP + FN)}$$

$$Specificity = \frac{TN}{(TN + FP)}$$

Results

Latent variable construction: The linear regression was used to test these latent variables. The R-square of mammography dataset is 0.83. It demonstrated that new latent variables are good consistent with other features. The R-square of clinical pathological dataset is 0.67. It demonstrated that new latent variables are fair consistent with other features. In addition, confirmation factor analysis was used for guarantee the compatibility the new feature with the empirical data. The statistical indicators test Randi (2012) shows the good result (see Table 1). Dataset fusion using Join Method: Integrated dataset samples were extended from hundred to 77,535. It found that not only incomplete and insufficient sample problems could be solved but also could be built early cancer staging, diagnosis, and prognosis model using the less aggressive method. The prediction models are divided into three parts. (1) Start with mammogram, staging could be predicted. (2) After by lymph node status for subtype staging follow by (3) Overall survival time is the last model.

Staging

The integrated dataset achieves to build the early staging prediction. The models accuracy is 98%. An area under ROC curve is 1. The rule inference mammogram to staging shows in Table 2. The rule sets are:

Stage I: The lower factors, tumor size (T1), shape Round=1, Density Low=1, Margin Well-defined=1, and BI-RADS 2-4 classify in Stage I.

Stage II: Margin in Obscured=2, Microlubulated=3, ill-defined=4, Density iso=3, BI-RADS5 are classify in Stage II.

Stage III: The high factor, Shape Lobular=3 or Irregular=4, Density iso=3 or high=4, Margin speculated=4, and BI-RADS4-5 classify in Stage III

However, the standard staging need the number of lymph nodes for stage grouping such as Stage IA, IB, IIA, IIB, IIIA, IIIB, etc. Then, we use integrated dataset to predict the number of lymph nodes.

Lymph node status

The early lymph node status prediction is important for rapid selecting the axillary lymph node dissection or prepared to sentinel lymph node biopsy. The lymph node status prediction performances are accuracy 72.47%, specificity 73.94%, and sensitivity 72.5%. An area under ROC curve is 0.74. The rules sets are:

N0/N1: Stage I or II and T1-T3 classify N0-N1 (0-3 nodes appeared).

N2: Stage III and T1-T3 classify N2 (4-9 nodes appeared)

N3: Otherwise in N3 (more than 9 nodes appeared).

These results show the lymph node status and tumor size. This information is useful for predicting the stage group including performances are accuracy 82.02%, specificity 87.9%, and sensitivity 82%. An area under ROC curve is 0.88. The rules sets are:

IA: Tumor size < 2 cm. and lymph node statue negative classify Stage IA.

IIA: Tumor size = T1, T2 and lymph node N0, N1 classify Stage IIA.

IIB: Tumor size = T2, T3 and lymph node N0, N1

Table 1. The Statistical Indicators Test

Indicator		Mammogram	Pathological
Chi-square	$0.05 \leq p \leq 1$	5.029, p = 0.540	8.361, p = 0.137
X^2/df	$0 < X^2/df \leq 2$	0.838	1.672
GFI	$0.95 \leq GFI \leq 1$	0.996	0.985
NFI	$0.95 \leq GFI \leq 1$	0.995	0.97
CFI	$0.97 \leq GFI \leq 1$	1	0.987
RMSA	$0 \leq RMSA \leq 0.05$	0	0.058

Table 2. The Rule Inference Mammogram to Staging

Staging	Inference Rules
Stage I	Tumor = T1 AND Margin = 1 AND Shape = 1 AND Density = 1: 1 (4417)
	Margin = 1 AND Shape = 1 AND Density = 2: 1 (5743)
	Margin = 1 AND Shape = 1 AND Density = 3 AND BIRADS = 4: 1 (5743)
	Margin = 1 AND BIRADS = 4 AND Density = 1: 1 (1325)
	Margin = 1 AND BIRADS = 3: 1 (5743)
	Density = 1 AND BIRADS = 2: 1 (2871)
Stage II	Margin = 2: 2 (2756)
	Margin = 3: 2 (5858)
	Margin = 5: 2 (4479)
	Shape = 3 AND Density = 3: 2 (3446)
	Density = 3: 2 (3446)
Staging	Inference Rules
	BIRADS = 5: 2 (2067)
	Else : 2
Stage III	Margin = 5 AND Shape = 3 AND Density = 3: 3 (3230)
	Margin = 5 AND Shape = 4 AND BIRADS = 5: 3 (8076)
	Margin = 5 AND Shape = 4 AND Density = 3 AND BIRADS = 4: 3 (4845)
	Margin = 4 AND Density = 4 AND BIRADS = 5: 3 (3230)
	Shape = 4 AND Density = 3 AND BIRADS = 5: 3 (4845)
	Shape = 4 AND Density = 4: 3 (1615)

classify Stage IIB.

IIIA: Tumor size = T3 and lymph node N1, N2 classify Stage IIIA.

IIIB: Tumor size = T4 classifies Stage IIIB.

IIIC: Lymph node N3 classifies Stage IIIC.

IV: Otherwise classify Stage IV.

The early breast cancer (IA, IIA) are T1-T2 and N0-N1. The locally advanced breast cancer (IIB, IIIA, IIIB, IIIC) is T2-T4 and N1-N3. Metastatic breast cancer is AnyT, AnyN, M1.

Overall survival time

The survival time defines as incidence of breast cancer where the cancer patient is alive less than one year, sixty months (5-years survival), or more than sixty months from the date of diagnosis. The model achieved to classify overall survival in three time periods. The OS prediction model performances are accuracy 72.72%, specificity 80%, and sensitivity 77%. An area under ROC curve is 0.87. The rules sets are:

OS < 1 year: This class shows the high risk factors Tumor size T2, T3, T4, Lymph node N1, N2, N3. The T1N2Mx, T2N2Mx, T3N1Mx, T3N2Mx are Stage IIIA. The T4N1Mx and T4N2Mx are Stage IIIB. The AnyT,N3Mx is Stage IIIC. The ER status probabilities are 49% positive and 51% negative. The HER2 status probabilities are 90% positive and 10% negative.

OS 1-5years: This class shows the factors Tumor size T2, T3, Lymph node N1, N2, N3. T2N3Mx is Stage IIIC. The T2N1Mx is Stage IIB and include mammogram factors Shape= oval, lobular or irregular, Margin more than obscured categories, Density=iso or high, BI-RADS4-5, and Age 51-75. The ER status probabilities are 38% positive and 62% negative. The HER2 status probabilities are 1% positive and 90% negative.

OS > 5 years: This class shows the low risk factors Tumor size less than 5 cm, Lymph node N0. The ER status probability is 100% positive and HER2 status 100% negative.

Performance comparison between single and integrated: The performance comparison between single clinical pathology and mammography dataset shows in Table 3. This work demonstrates that the rule induction model may success for rapid breast cancer screening, diagnosis, and prognosis.

The results are not different form the standard methods. However, some factors are not appearing in mammography screening. The integrated dataset achieves to complement dataset and more improve accuracy than single dataset.

Table 3. Comparison between Single Clinical Pathology and Mammography Dataset

Model	Accuracy (Single)	The number of rules (Single)	Accuracy (Integrated)	The number of rules (Integrated)
Mammogram to Staging	68.93	8	100	20
Lymph node	74.75	9	72.47	10
Overall survival time	45.7	8	76.54	26

Discussion

The staging prediction over mammogram screening: the results from mammography represent the tumor characteristic and classify in BI-RADS1-5. Normal mammograms classify an assessment category of "negative" or "benign finding" (Berta et al., 2001). The women in the screening study who were found to have a category 4 "suspicious abnormality" were much less likely to have a recommendation of clinical consultation or biopsy as well as category 3 (probably benign) mammograms showed similar variation in the association to management recommendation. Our study extended findings not only classify BI-RADS assessment but also classify early staging in Stage I, II, III, and IV when patients who receives mammography. The models represent in easy decision rules. The low value factors show the lower stage than high value such as Sate I get BI-RADS 2-4 and Stage III get BI-RADS 5. Some of the aims for staging are 1) Aid medical staff in staging the tumor helping to plan the treatment 2) Give an indication of prognosis and 3) Assist in the evaluation of the results of treatment.

The lymph node status prediction over mammogram screening: The staging prediction models from mammography have been predicted the cancer staging that is the first screening in breast cancer. However, the standard staging need the lymph node status for stage grouping such as Stage IA, IB, IIA, IIB, IIIA, IIIB, etc. Only mammography is not report the lymph node status. Clinically detected is defined as detected by clinical examination or by imaging studies (excluding lymphoscintigraphy) and having characteristics highly suspicious for malignancy or a presumed pathological macrometastasis based on fine-needle aspiration biopsy with cytological examination (Sobin and Wittekind, 2009). So, the stage group classification needs the model to predict lymph nodes. A computer-aided prediction, it has been reported (Woo et al., 2017) hypothesized that breast tumor surrounding tissue features in ultrasound images might be useful to predict axillary lymph nodes in breast cancer. Their results showed the textural feature set has higher performance than intensity feature set and morphology feature set in prediction of axillary lymph nodes. Other studies (Pinheiro et al., 2014; Fidan et al., 2016) axillary lymph nodes have been predicted by primary tumor features, sonographic characteristics of lymph node and clinical data. Combine dataset approach (Marco et al., 2016) concluded that combined axillary ultrasound and FNA biopsy had a high accuracy rate for the preoperative diagnosis of the axilla. Our study also had finding similar to those of study by the integrated dataset. The lymph node prediction model has been predicted the number of lymph nodes for stage group classification. The model shows the rule inference to lymph node N0, N1, N2, and N3 by head staging I, II, and III. When we know the number of lymph node, the lymph nodes and tumor size can be built the stage group such as IA, IB, IIA, IIB, IIIA, IIIB, etc. The rules are similar with TNM staging system but the challenge is the integrated data extended mammography information to early diagnosis. In addition, the advantage of early diagnosis, staging, and lymph node prediction, help the physician to plan the best treatment approaches such as selecting metastases patients eligible to immediate ALND without a preliminary SLNB for the pre-operative axillary staging as well as for selecting patients eligible to the rapid treatment.

The prognosis model: Breast cancers can be classified by different factor. Each of these aspects influences treatment response and prognosis. Description of a breast cancer would optimally include all of these classification aspects, as well as other findings, such as signs found on physical exam. A full classification includes histopathological type, grade, stage (TNM), receptor status, and the presence or absence of genes as determined by DNA testing. Prognosis factors (Cianfrocca and Goldstein, 2004) are any measurement available at the time of surgery that correlates with disease-free or overall survival such as tumor size, axillary node status, and tumor grade. The most informative features used for prognosis in previous work (Kanghee et al., 2013) are the tumor size, the number of nodes and the age at diagnosis. In integrated dataset not only use the similar prognostic factor but also use mammography factor to build the prognosis models. The overall survival times was divided in three categories (OS<1 years, OS 1-5 years, and OS > 5 years). The our experiments show the rules in the high risk factors, Stage IIIA, Stage IIIB, Stage IIIC, ER status negative, and HER2 positive, is poor prognosis (OS<5 years). The rules in low risk factors, tumor size less than 5 cm, lymph node N0, ER status positive, and HER2 status negative, is good prognosis (OS>5 years).

The report from The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute based on data from SEER 18 2007-2013 shows the 5-year survival for distance female breast cancer is 6%. The 5-year survival for regional female breast cancer is 31% and localized female breast cancer is 62%. Integrated dataset shows the 5-year survival for distance female breast cancer is 1%. The 5-year survival for regional female breast cancer is 42% and localized female breast cancer is 57%. Our finding, the 5-year survival rate by extended prognosis factor not only anatomic staging but also extended mammography, ER status, and HER2 status, have been reduced 5-year survival rate, 5% in distance and localized female breast cancer because the extended factors (mammography, ER status and HER2 status) affected with the treatment outcome. However, the 5-year survival rate for regional female breast cancer is 11% increased because the extended factors affected with the best treatment approach. The Breast cancer screening and diagnosis guideline (Chaiveerawattana et al., 2012) recommend axillary lymph node dissection when the number of lymph nodes are more than ten nodes. Other management used sentinel lymph node biopsy instead. Our experiment show 98% of surveyed would avoid the axillary lymph node dissection and 2% would not delay to rapid axillary lymph node dissection.

This work aims to build the prediction model for rapid breast cancer screening, diagnosis, and prognosis overall survival time using integrated dataset that combined two datasets by latent variables. The contributions this work

are (1) predict the early cancer stage over mammography (2) estimate the lymph node for selecting metastases patients to ALND or SLNB (3) evaluate overall survival time. The results are not different form the standard prognosis. However, some factors are not appearing in mammography screening. The integrated dataset mammography and clinical pathological extended factors for decision better than using single dataset. Furthermore, integrated dataset improve model performance compared with single dataset. Other factors are important to predict the survival time such as other biomarker, surgery operation, neoadjuvant therapy, hormone-therapy, or chemotherapy will be included the prediction models.

References

Ahmad A, Negm S (2015). Robust breast cancer diagnosis on four different datasets using multi- classifiers fusion. *Int J Eng Res Technol Res*, **4**, 114-8.

Berta M, William E, Virginia L (2002). Use of the American College of Radiology BI-RADS to Report on the Mammographic Evaluation of Women with Signs and Symptoms of Breast Disease. *Radiogy*, **222**, 536-42.

Chaiveerawattana A, Sukyothin S, Imsamran w (2012). Breast cancer screening and diagnosis guideline. National Cancer Institute, Thailand.

Cianfrocca M, Goldstein L (2004). Prognostic and predictive factors in early-stage breast cancer. *Oncllogist*, **9**, 606-16.

Pinheiro DJPC, Elias S, Nazário ACP (2014). Axillary lymph nodes in breast cancer patients: sonographic evaluation. *Radiol Bras*, **47**, 240-4.

Gouda I, Abdelhalim MB, Magdy Z (2012). Breast cancer diagnosis on three different datasets using multi-classifiers. *Int J Comput Inf Technol*, **1**, 36-43

Jingming Ye, Wenjun W, Ling Xu, et al (2017). A retrospective prognosis evaluation analysis using the 8th edition of American Joint Committee on Cancer (AJCC) cancer staging system for luminal A Breast cancer. *Chin J Cancer Res*, **29**, 351-360.

Konstantina K, Themis P, Konstantinos P (2015). Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*, **13**, 8-17.

Kanghee P, Amma A, Dokyoon K et al (2013). Robust predictive model for evaluating breast cancer survivability. *Eng Appl Artif Intell*, **26**, 2194-2205.

Linqi S, William H, Jie Xu, et al (2016). Using contextual learning to improve diagnostic accuracy: Application in Breast Cancer Screening. *IEEE J Biomed Health Inform*, **20**, 902-14.

Marco G, Piero F, Alessandro G (2016). Axillary ultrasound and Fine-Needle Aspiration Cytology in the preoperative staging of axillary node metastasis in breast cancer patients. *Breast J*, **30**, 146-50

Ogunsakin R, Siaka L (2017). Bayesian inference on malignant breast cancer in Nigeria: A Diagnosis of MCMC Convergence. *Asian Pac J Cancer Prev*, **18**, 2709-16.

Randi H, Olsson U (2012). Testing structural equation models: The impact of error variances in the data generating process. *Qual Quant*, **46**, 1547–70.

Ralph L (2014). Latent Variables. Encyclopedia of Quality of Life and Well-Being Research, pp 3523.

RapidMiner GmbH (2017). RapidMiner Studio Manual. Global leader in predictive analytics software.

Woo M, Yan-Wei L, Yao-Sian H (2017). Computer-aided prediction of axillary lymph node status in breast cancer using tumor surrounding tissue features in ultrasound images. *Comput Methods Programs Biomed*, **146**, 143-50.

Yeye H, Kris G, Xu C (2015). SEMAJOIN: Joining Semantically Related Tables Using Big Table Corpora. Proceedings of the VLDB Endowment, **8**, pp 1358-69.