Validation of a nomogram to predict axillary lymph node status in Dutch early breast cancer patients

Student

M. Aarnink

S1114972

Date August 2016

integraal kankercentrum Nederland

UNIVERSITY OF TWENTE.

Validation of a nomogram to predict axillary lymph node status in Dutch early breast cancer patients

Date of submission	23-08-2016
Date of colloquium	30-08-2016

Surname	Aarnink
Given name	Merel
E-mail	merelaarnink@gmail.com

Master programme	Health Sciences
Track	Health Services and Management
Faculty	Faculty of Science and Technology

First supervisor	Prof. Dr. S. Siesling, University of Twente, Netherlands Comprehensive Cancer Organisation (IKNL)
Second supervisor	Dr. H. Koffijberg, University of Twente
External supervisors	M.C. van Maaren, Msc, Netherlands Comprehensive Cancer Organisation (IKNL) Prof. Dr. G.M. van Dam, Universitair Medisch Centrum Groningen S. Qiu, Msc, Universitair Medisch Centrum Groningen, Cancer Hospital of Shantou University Medical College

Preface

This study is the result of my master assignment of Health Sciences at University of Twente. I started working on this assignment in February 2016. I was really enthusiastic after reading the goal of this assignment for several reasons. First, I was looking for a more quantitative research, since I find this more fun and challenging. Second, the study was focussed on breast cancer, which I find a really interesting and socially relevant subject. Last, I always had “something” with technologies. A great part of this study was focused on ultrasound and ultrasound images. It was a nice challenge for me to test and improve my knowledge on this diagnostic tool.

During this assignment I have learned many things. First, off course, I improved my knowledge on breast cancer, the axillary lymph node status, the role of ultrasound in staging these nodes and on the use and working of predictive models. Second, I have learned a lot about the research process in practice. During my study, assignments were very organised and this assignment showed me how things are (a lot) more unorganized in practice. While writing my research proposal it was of great importance to get permission in hospital for data collection and during data collection I had to start with the data analyses on the first subpopulation. Research is not a step-wise process in practise, which I knew but had to experience. Lastly, I learned a lot about myself and this research made me more confident about my knowledge and ability.

This thesis was performed internally at the Comprehensive Cancer Organisation (IKNL) and data was collected in six different hospitals. Many people contributed to this study in some way. First, I would like to thank my first and second supervisors, Sabine Siesling and Erik Koffijberg, for guiding me through this assignment. I have learned a lot in our meetings and your feedback was of great value for this final version. I also want to thank Gooitzen van Dam and Siqi Qui, for making this assignment possible. Siqi, it was nice having someone performing the same analyses as back up. Our, sometimes, differing view on things helped me making this thesis better. Third, I want to give special thanks to Marissa van Maaren, my external supervisor from IKNL. Our weekly meetings and your feedback kept me going on track and kept me positive all time. Last but not least, I want to thank all participating radiologists, Monique Dorrius, Jan Korte, Jeroen Veltman, Caroline Klazen, Pieter Ott and Susanne Estourgie (surgeon). All of you contributed to the data collection and I could not have done it without your expertise.

Merel

Enschede, August 2016

Abstract

Introduction: Axillary lymph node (ALN) status is a an important prognostic factor in patients with primary breast cancer. Sentinel lymph node biopsy (SLNB) is standard procedure for staging the axillary lymph nodes. However, the procedure is related to some risks on morbidity. To avoid axillary surgery in patients with low risk on metastases, a nomogram to predict ALN status based on a Chinese population is developed in 2016. The model showed good performance in an internal validation population. The aim of this study was to externally validate this model in a Dutch population.

Methods: Early stage breast cancer patients from six Dutch hospitals, diagnosed between January 2011 and December 2015, and with positive ultrasound findings were included. Patients who received primary systemic therapy, patients with bilateral breast cancer or patients with incomplete data on the variables in the nomogram were excluded. The validation population was compared to the development population. The area under the receiver operating characteristic (ROC) curve (AUC), false negative rates (FNR) and false omission rates (FOR) were calculated to determine the predictive accuracy. The Hosmer-Lemeshow goodness of fit (HL) test and a calibration plot were used to assess its goodness of fit. The model was updated using logistic regression.

Results: 1,416 patients were included in this validation study. Large differences in tumour- and ALN characteristics were found in the development- and validation population. Only transverse diameter showed no significant difference. 24.93% of all patients in the validation population had ALN metastases, compared to 50.62% of the patients in the development population. The AUC was 0.77. The HL- test showed a significant difference between the predicted probability and the observed event rate. However, the calibration plot showed a good fit. In the updated model, the FNR was 5.67% at a cut-off point of 7%.

Discussion: Based on predictive accuracy and calibration, the model seems to perform good in the Dutch patient population. The updated model has better predictive accuracy for selecting low-risk patients when compared to the original model. Although the model cannot serve as preoperative information tool in the Netherlands, the model can be of possible value in other Western European countries. For the Dutch situation, it would be worthwhile to adapt the model or to develop a new model with preoperative variables only.

Keywords: breast cancer; axillary lymph node status; ultrasound; predictive model; validation

Table of contents

1 Introduction

The axillary lymph node (ALN) status is one of the important prognostic determinants in patients with primary breast cancer. (1, 2) Ultrasound is considered to be one of the most significant preoperative imaging tools for evaluation of the ALNs. (3-6) A systematic review of Alvarez et al. showed that sensitivity, however, was moderate and varied between 26.4% (95% CI 15.3 – 40.3%) and 75.9% (95% CI 56.4 – 89.7%). (4) Ultrasound has the best sensitivity in lymph nodes that appear abnormal. (4, 7) In normal-appearing lymph nodes, ultrasound does not provide sufficient information for good classification of the ALNs. (8) For those patients, additional treatment is needed.

In the Netherlands, ultrasound of the axillary area is performed to determine the clinical ALN status in women with breast cancer. (9) The cortical thickness and the absence of a hilum are two of the most important predictors for ALN metastases. (10, 11) For the cortical thickness a cut-off point of 2.3 mm is used. (12) A fine needle aspiration (FNA) is recommended if a suspicious lymph node is found. (12) Based on ultrasound or FNA, the physician will decide what axillary treatment a patient receives for determining the pathological ALN status. Nowadays, sentinel lymph node biopsy (SLNB) is, as in many other countries, indicated for Dutch patients with T1-2N0 breast cancer. (9, 13) ALN dissection (ALND) is often reserved for patients with lymph node metastases proven by a positive fine needle aspiration or SLNB. (9, 14) The physician can also decide to start with axillary radiotherapy. (14) A flowchart of the decision process for axillary surgery is shown in *Figure 1*.

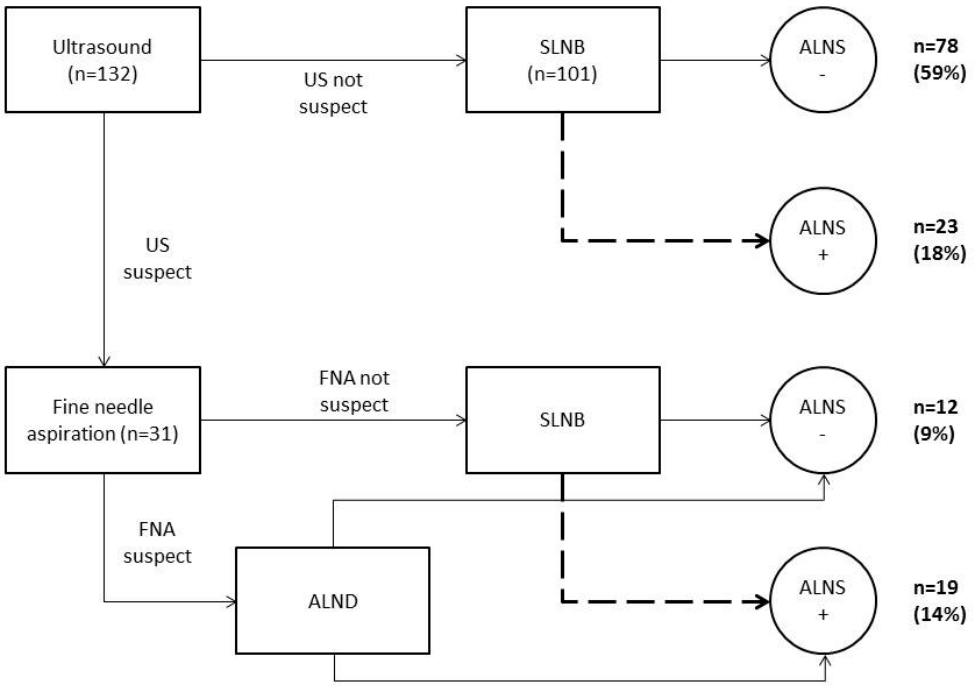


Figure 1 Flowchart of axillary treatment, based on Nori et al. (2007)(15)
Note: ALNS: axillary lymph node status; ALND: axillary lymph node dissection; SLNB: sentinel lymph node biopsy; FNA: fine needle aspiration

ALND and SLNB are both associated with disadvantages like lymphedema, limited arm and shoulder movement and numbness. (16-19) Although SLNB has significant less side effects and offers outcomes comparable to ALND, the risk of morbidity can not completely be eliminated. (17-21) Due to improved screening techniques, the incidence of early stage breast cancer has increased and the presence of ALN metastases has declined. (22) This results in more patients for whom axillary surgery and its possible side-effects could have been omitted. Results of a study of Nori et al. are summarized in *Figure 1*. As can be seen, axillary surgery could have been omitted in 90 patients (68%) without metastases. Also, between 23 and 42 patients (18 – 32%) with metastases received both SLNB and ALND (dashed line), instead of having them directly receiving the right treatment. (15)

Currently, studies are looking into the possibilities to safely omit ALND in breast cancer patients. (14, 23) In 2014, a trial demonstrated that radiotherapy could be a safe alternative to ALND in patients with sentinel lymph node metastasis. (14) According to Galimberti et al. ALND could be avoided in patients with early stage breast cancer. (23) These results lead to doubts about the role of SLNB itself, since SLNB is mainly used in decision-making for further axillary treatment. (24) Considering the disadvantages of these ALN treatments and to supress overtreatment, it is of great clinical value to predict the ALN status preoperatively whereby axillary surgery can be omitted in low risk patients.

Different predictive models are developed to predict ALN status in breast cancer patients. Some models tend to predict the general ALN status in proven sentinel lymph node positive patients. (25, 26) However, it would be of more value if the ALN status could be determined preoperatively. Currently, models with good accuracies (AUC 0.731 (27) and 0.849 (28)) are developed. However, these models cannot be used preoperatively, do not include ultrasound variables and/or they are based on a small patient population. (27, 28)

In 2016, a nomogram to predict the ALN status was developed on a Chinese patient population. The aim of the model was to provide a preoperative tool for assisting clinical decision-making. The model includes patients with at least one lymph node detected on ultrasound and tends to predict the overall ALN status. The probability on ALN metastases is calculated based on six tumour and axillary ultrasound related variables. The model already showed good discrimination (AUC 0.864) in an internal validation population. (29) For further generalizability and to evaluate the predictive ability, the model needs to be validated in external validation groups. The aim of this study was to validate the nomogram for predicting the probability of ALN involvement in Dutch breast cancer patients and to assess the possible clinical value for the Dutch population.

2 Methods

Setting and subjects

For this validation study, a population of 2,940 women with early stage invasive breast cancer (C50, T1-3, N0-1) and known pathological lymph node status was selected from the Netherlands Cancer Registry (NCR). The NCR, managed by IKNL, includes data related to patient, tumour and treatment characteristics of all cancer patients in the Netherlands. (30) For this study, patients with positive ultrasound findings diagnosed in six Dutch hospitals between January 2011 and December 2015 were included. A positive ultrasound was defined as the detection of one or more lymph node(s). Participating hospitals were Medisch Spectrum Twente, Ziekenhuisgroep Twente, Isala, Universitair Medisch Centrum Groningen, Medisch Centrum Leeuwarden and Martini Ziekenhuis. Patients who received primary systemic therapy or patients with bilateral breast cancer were excluded. Breast cancer was considered bilateral if a tumour in both breasts was found at time of diagnosis. Patients with incomplete data on any of the six variables in the model were also excluded for validation.

Data collection

Patient-, tumour- and ALN characteristics were gathered from the NCR and patient radiology reports. Age at time of diagnosis, menopausal status, three-numbered topography code (localisation and sub localisation), histological grade, oestrogen receptor (ER) status, progesterone receptor (PR) status and human epidermal growth factor receptor 2 (HER2) status were gathered from the NCR. The clinical tumour size and characteristics of the lymph node were additionally collected from ultrasound images and patient reports. Characteristics of the lymph node included transverse diameter, cortical thickness and presence/absence of hilum.

Radiology reports and ultrasound images of 2,227 patients were checked by one author (MA) after receiving a short training. The patients were randomly ordered and there were as many patient reports checked as possible. The patient was excluded if there were no images of the axillary area present in the radiology reports, or if the absence of lymph nodes on the ultrasound was obvious. If there were lymph nodes detected and measured on ultrasound, or if they were relatively easy to measure, the patient was included in the study population. If there were any doubts about the presence or measures of the lymph nodes, the ultrasound images were reassessed by an experienced radiologist of the concerning hospital. At the end of the inclusion phase, the validation population consisted of 1,416 patients, who met the in- and exclusion criteria. *Figure 2* gives an overview of this patient selection.

Definitions

Topography and morphology were coded according to the International Classification of Diseases for Oncology (ICD-O). (31) Staging was coded according to the Tumour, Node and Metastasis (TNM) classification system. (32) In this study, HER2, ER- and PR status are categorized in the same way as in the article of Qiu et al. Both variables are divided into four subcategories: ER/PR– (<10%); ER/PR + (10-25%); ER/PR++ (25-75%) and ER/PR +++ (>75%). (29) Since the NCR registers the ER- and PR

status as 0, 10%, 20%, 30%, etc., it was chosen to replace the second cut-off point (25%) with 30% and the third cut-off point (75%) with 80%. The coding of HER2 was done according to the Dutch clinical guideline for breast cancer. (9) The Molecular subtype is based on the HER2 and hormone receptor statuses. The variables were measured on the most suspected lymph node according to recent literature, in this study considered as the lymph node with the thickest cortex and/or with an absence of hilum.

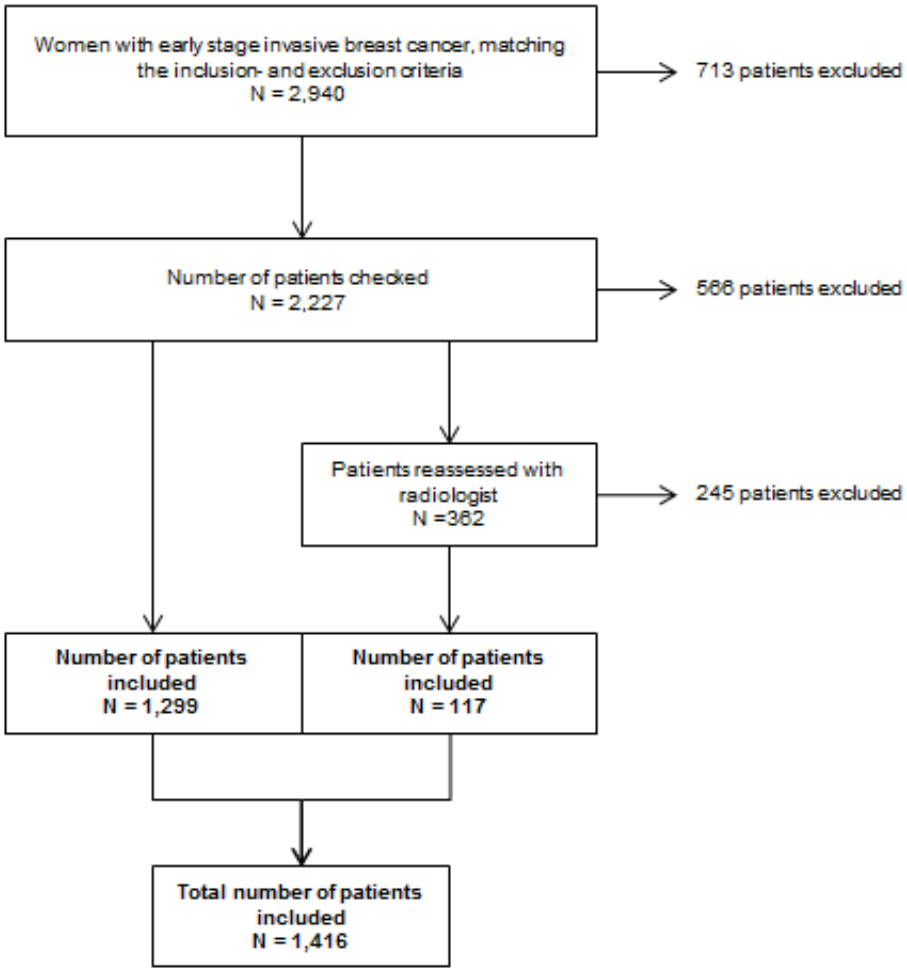


Figure 2 Flowchart: research population

Statistical methods

To compare the model development population with the validation population, patient-, tumour- and ALN characteristics were compared using the Chi^2 test for categorical variables and the Mann-Whitney U test for continuous variables.

The predictive accuracy of the model was assessed using the area under the receiver operating characteristic (ROC) curve (AUC), from now abbreviated to AUC. Since there is no agreement on what AUC-value represents a good quality, the estimated accuracy was compared with AUC-values of other predictive models regarding ALN status. (33) The false negative rate (FNR) and the false

omission rate (FOR) were additionally calculated to assess the predictive accuracy. These were compared to FNRs and FORs in other studies, in particular with the internal validation study. The Hosmer-Lemeshow goodness of fit test (HL-test) was used to evaluate the goodness of fit of the model. Hereby, the validation population was divided into deciles of predicted risk, after which the average predicted risk was compared to the observed event rate, separately for each group. A good fit is assumed with a p-value above 0.05. Additional information about the goodness of fit is gathered from a calibration plot. Sensitivity, specificity, FNR and FOR were calculated for different threshold values up to 20%, based on the threshold values chosen in the article of Qiu et al. to make a good comparison. (29) Based on these threshold values, it is intended to select patients with low risk on having ALN metastases for omission of axillary surgery.

Updating the model was considered by adjusting the intercept and slope of the original model to improve performance in the validation population. The linear predictor was separated from the original model:

$$\begin{aligned}\text{Linear predictor} = {} & 0.063 * \text{transverse diameter} + 0.277 * \text{cortical thickness} + \\ & 1.420 * \text{hilum absent} + 1.502 * \text{histological grade 2} + 2.090 * \text{histological grade 3} + \\ & 0.305 * \text{clinical tumour size} + 0.379 * \text{ER status})\end{aligned}$$

The correction factor for slope and the new intercept were determined using logistic regression analysis, with axillary metastases as dependent variable and the linear predictor as independent variable. To assess the models' improvement, the predictive accuracy of the updated model at different threshold values was compared with the predictive accuracy in the original model. Also, the HL-test was again performed to see if the updated model was a better fit for the study population, compared with the original model.

Stata/SE 14.1 was used for all analyses and a p-value of 5% was considered as statistically significant.

Ethical consideration

This research was based on existing patient data, and did not subject individuals to any intervention. Therefore this research was not covered by the Law Medical Research (Wet Medisch Wetenschappelijk onderzoek). (34) The results are presented only on aggregated level, so it is not possible to trace information on patient level.

3 Results

Patient characteristics

An overview of patients characteristics in the development population (n=322) and validation population (n=1,416) is given in *Table 1*. Only transverse diameter does not show a significant difference between the Chinese development and Dutch validation population (p = 0.70). A large difference can be seen in clinical tumour size (in mm and as TNM stage). The median of the clinical tumour size in the validation population (15mm, 95% CI 10 – 22) is only half of the median of the clinical tumour size in the development population (30mm, 95% CI 23 – 40). This is also the case for cortical thickness, for which the median is 1.9 mm (95% CI 1.3 – 2.8) in the validation population and 4 mm (95% CI 3 – 6) in the development population. Another large difference can be seen in the amount of patients with absence of hilum in the lymph node. The hilum is absent in a larger proportion of patients in the development population (39.13%) than in the validation population (7.98%). Finally, a notable difference is seen in the number of patients with pathological proven lymph nodes. There are more than twice as much patients diagnosed with ALN metastases in the development population (50.62%) as in the validation population (24.93%).

Table 1 Comparing development population and validation population by patient-, tumour- and ALN characteristics

Variable	**Development population No. (%)** ***n=322***	**Validation population No. (%)** ***n=1,416***	**p-value**
Region	China, Guangdong	The Netherlands	
Age at diagnosis – Median (IQR)	50 (43, 57)	61 (52, 69)	**<0.001**
Menopausal status			**<0.001**
Premenopausal	182 (56.52)	201 (14.19)	
Postmenopausal	140 (43.48)	1,069 (75.49)	
Perimenopausal	-	73 (5.16%)	
Unknown	-	73 (5.16%)	
Clinical tumour size (mm) – Median (IQR)	30 (23, 40)	15 (10, 22)	**<0.001**
Clinical tumour size (TNM)			**<0.001**
T1	74 (22.98)	926 (65.40)	
T2	223 (69.25)	459 (32.42)	
T3	22 (6.83)	31 (2.19)	
Unknown	3 (0.93)	-	
Tumour location			**<0.001**
UOQ	152 (47.20)	573 (40.47)	
LOQ	42 (13.04)	118 (8.33)	
UIQ	51 (15.84)	178 (12.57)	
LIQ	15 (4.66)	127 (8.97)	
Central	62 (19.25)	102 (7.20)	
Overlapping lesions	-	309 (21.82)	
Unknown	-	9 (0.64)	

Variable	Development population No. (%) *n=322*	Validation population No. (%) *n=1,416*	p-value
Histological grade			**<0.001**
I	49 (15.22)	375 (26.48)	
II	104 (32.30)	676 (47.74)	
III	154 (47.83)	365 (25.78)	
Unknown	15 (4.66)	-	
ER status			**<0.001**
Negative	119 (36.96)	229 (16.17)	
1+	22 (6.83)	25 (1.77)	
2+	57 (17.70)	81 (5.72)	
3+	124 (38.51)	1,081 (76.34)	
PR status			**<0.001**
Negative	132 (40.99)	388 (27.40)	
1+	38 (11.80)	105 (7.42)	
2+	63 (19.57)	242 (17.09)	
3+	89 (27.64)	681 (48.09)	
Her-2 status			**<0.001**
Negative	223 (69.25)	1,242 (87.71)	
Positive	99 (30.75)	162 (11.44)	
Unknown		12 (0.85)	
Molecular subtype			**<0.001**
Luminal A	174 (54.04)	1,066 (75.28)	
Luminal B	43 (13.35)	116 (8.19)	
Her-2 enriched	56 (17.39)	46 (3.25)	
Triple negative	49 (15.22)	176 (12.43)	
Unknown	-	12 (0.85)	
Transverse diameter (mm) –Median (IQR)	13 (10, 17)	12.6 (9.6, 16.9)	**0.70**
Cortical thickness (mm) – Median (IQR)	4 (3, 6)	1.9 (1.3, 2.8)	**<0.001**
Absence of hilum			**<0.001**
Yes	126 (39.13)	113 (7.98)	
No	196 (60.87)	1,303 (92.02)	
ALN metastases			**<0.001**
Yes	163 (50.62)	353 (24.93)	
No	159 (49.38)	1,063 (75.07)	

Note: UOQ: upper outer quadrant; LOQ: lower outer quadrant; UIQ: upper inner quadrant; LIQ: lower inner quadrant; IQR: interquartile range; ER: oestrogen receptor; PR: progesterone receptor; HER-2: human epidermal growth factor receptor 2; ALN: axillary lymph node.

Model validation

The AUC (95% CI) is 0.77 (0.74 – 0.80) for the validation population (*Figure 3*), which is lower than the AUC derived from internal validation in the development population (AUC = 0.86). The HL-test showed a significant difference between the predicted and observed probability of ALN metastases (p<0.001). *Figure 4* shows the calibration plot of the model in the validation population. The model gives an overestimation of the probability on having metastases in decile 4, 7, 8 and 9, where especially the consecutive deciles (7-9) are notable. A relatively large underestimation can be seen for the second

decile. The calibration line and the reference line lie very close to each other, both in the 95% confidence interval. *Table 2* shows the sensitivity, specificity, FNR and FOR at different threshold values. At a threshold value of 6%, 4.82% of all patients with pathologically proven metastases had a negative test outcome. At a threshold value of 8% the FNR had almost doubled to 9.07%. The FOR is 7.20% at a threshold value of 7% and 10.22% at a threshold value of 8%.

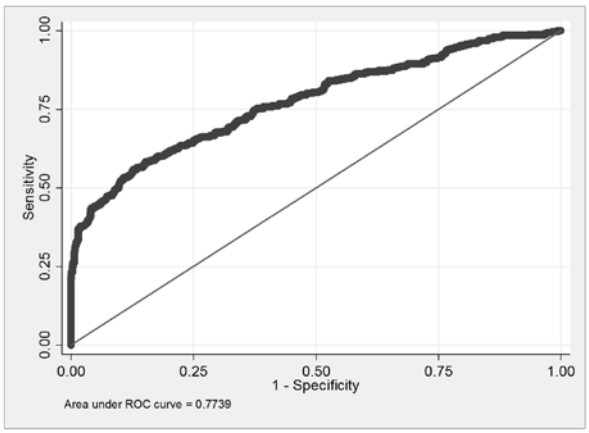


Figure 3 Area under the receiver operator characteristic curve

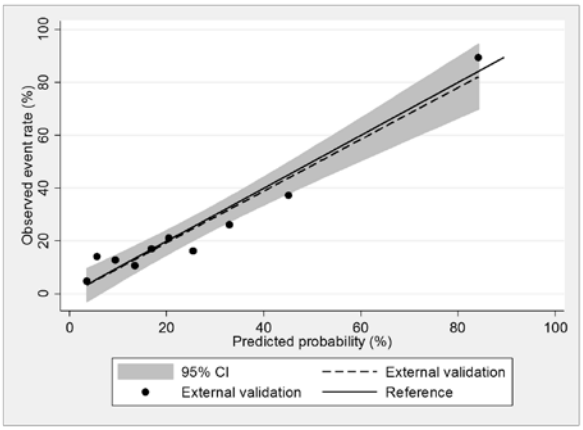


Figure 4 Calibration of the model in the validation population

Model update

Since the predicted probability on having ALN metastases differed significantly from the observed probability, it was decided to update the slope and intercept of the original model. The new intercept was – 5.31 and the correction factor for the linear predictor (lp) was 0.90. Again, the HL-test was used to assess the goodness of fit for the validation population. A significant difference between observed and predicted probabilities was found ($p<0.001$). A calibration plot of both the updated and the original model is shown in *Figure 5.* The differences between the original and the updated model look very small, in particular in deciles with lower predicted probabilities. Up to a predicted probability of around 10%, both the original and the updated model show an underestimation of the probability on ALN metastases (*Figure 5)*. In this range, the predicted probability of the updated model is higher than the original model. Sensitivity, specificity and FNR for the updated model are presented in *Table 2.*Up to a threshold value of 10%, the FNR of the updated model is lower than the FNR of the original model.

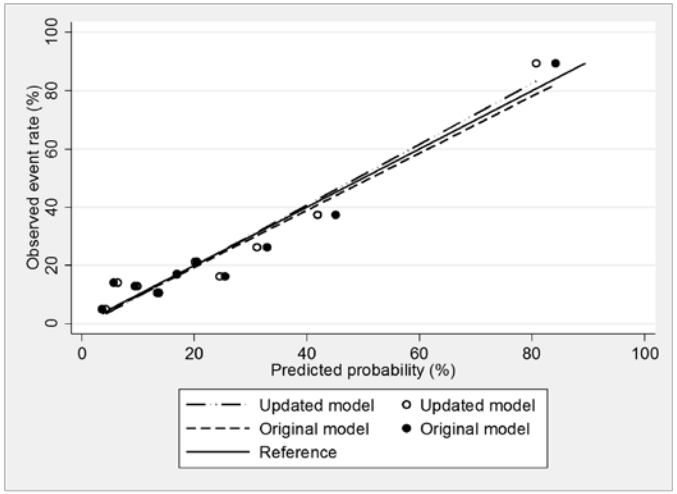


Figure 5 Calibration of the original- and updated model in the validation population

Table 2 Sensitivity, specificity and false negative rate of the original model and the updated model in the validation population at different threshold values

Threshold value	No. of patients beneath threshold (%)		No. of ALN metastases beneath threshold		Sensitivity (%)		Specificity (%)		FNR			FOR		
	Original model	*Updated model*	*Original model*	*Updated model*	*Original model*	*Updated model*	*Original model*	*Updated model*	*Original model*	*Updated model*	*Qui et al. (2016)*	*Original model*	*Updated model*	*Qui et al. (2016)*
<5%	168	113	10	5	97.17	98.58	14.86	10.16	2.83	1.42	-	5.95	4.42	-
<6%	236	198	17	14	95.18	96.03	20.60	17.31	4.82	3.96	-	7.20	7.07	-
<7%	286	250	27	20	92.35	94.33	24.37	21.64	7.65	5.67	-	9.44	8.00	-
< 7.1%	289	258	30	21	91.50	94.05	24.37	22.30	8.50	5.95	0.00	10.38	8.14	0.00
<8%	313	296	32	30	90.93	91.50	26.43	25.02	9.07	8.50	-	10.22	10.14	-
<9%	336	319	37	32	89.52	90.93	28.13	27.00	10.48	9.07	-	11.01	10.03	-
<10%	369	355	37	37	89.52	89.52	31.23	29.92	10.48	10.48	-	10.03	10.42	-
<13.8 %	510	498	53	51	85.00	85.55	43.00	42.05	15.00	14.45	0.75	10.39	10.24	5.26
<18.2 %	689	693	82	82	76.77	76.77	57.10	57.48	23.23	23.23	1.50	11.90	11.83	8.70
<20 %	761	774	93	96	73.65	72.80	62.84	63.78	26.35	27.20	3.76	12.22	12.40	17.24

Note: ALN: axillary lymph node; FNR = false negative rate; FOR = false omission rate

4 Discussion

Main findings

The aim of this study was to validate the predictive model of Qui et al. for ALN status in breast cancer patients on an external Dutch patient population and to assess the possible value in Dutch clinical practice. Based on discrimination and calibration, the model seems to perform good in the Dutch patient population and based on this performance the model could be potentially be used in future clinical practice after taken into account and removing the limitations of this model.

All variables differed significantly between the validation population and the development population, except for transverse diameter. The most remarkable differences were found in the number of patients with ALN metastases and in clinical tumour size (in mm and as TNM stage). The Dutch population consist mostly of T1 stage breast cancer patients (65.40%) and the Chinese population consist mostly of T2 stage breast cancer patients (69.25%). This difference between China and a western country is also found in another study. (35) A possible explanation for these differences is that Chinese women are diagnosed later. The absence of a nationwide screening programme that is reimbursed by the government is a possible reason for later diagnosis in China.(36) In the Netherlands, the screening programme led to an earlier diagnosis, leading to an increase of early stage breast cancer (T1N0). (22) Second, as confirmed in this study, the age of breast cancer patients is lower in China compared to western countries.(35, 37) This in combination with the reduced accuracy of mammography in younger patients could also lead later diagnosis. (38)

Despite the differences between the development population and the validation population, the model still shows good predictive accuracy with an AUC of 0.77, indicating good generalizability. The HL-test showed a significant difference between the observed probability and the predicted probability, which implies the model is not a good fit for the Dutch patient population. However, it is common that a statistical test tends to be significant if the study population is large enough. When validating a model developed on smaller patient population, in a larger number of patients, a significant HL-test does not mean that the model is not useful. (39) Based on this information it is chosen to look at the calibration plot instead of the p-value of the HL-test. When looking at the calibration plot in *Figure 3*, the line drawn through the scatter plot of the validation population almost overlaps the reference line. Based on the calibration plot, it is believed that the model gives a good calibration for the Dutch patient population.

Despite the good performance of the original model in the Dutch population, it was decided to update the model to see if a better fit of the model was possible. Since we want to select low patients for omission of axillary surgery, the lower probabilities should be more accurate. Up to a predicted probability of 10%, the underestimation of the updated model is less than the original model. To use the model in clinical practice, a threshold value should be chosen to select low-risk patients. *Table 2* shows predictive accuracy at different threshold values. It can be seen that the updated model,

indeed, performs better at low threshold values than the original model. The right threshold value was chosen based mostly on FNR. Since the FNR of SLNB varies between 5-10%, a FNR around 5% was considered acceptable. (40-42) At a threshold value of 7%, the FNR is 5.67% which would still be acceptable. At this threshold value, axillary surgery could be omitted in 21.64% of all patients without lymph node metastases. This suggests that, despite the underestimation in low predicted probabilities seen in *Figure 5*, the model still performs good enough.

Strengths and limitations

There are some limitations related with the model itself. First, ER status is treated as a continuous variable, where it is actually a categorical variable. It is not known if ER status would still be a predictor if it was included as categorical variable. Second, in this predictive model, transverse diameter is a predictor for getting ALN metastases. Since the development population consists of more patients with metastases, a larger median of transverse diameter is assumed for these patients. However, transverse diameter shows no significant difference between the two populations, so the importance of this variable for the probability on having ALN metastases is questionable. Last, the model was developed to serve as preoperative predictive tool. However, in the Dutch clinical situation, ER-receptor status and histological grade are based on primary surgery of the tumour. Nowadays axillary surgery is often performed at the same time as primary surgery. By splitting these surgeries, axillary surgery can be avoided in low-risk patients. However, all other patients will need surgery twice, which enlarges the burden on the patient.

This validation study is also related with some strengths and limitations. This study is performed on a large patient population (1,416), whereby 353 patients had ALN metastases. This research included patients from six different Dutch hospitals and every hospital employed several radiologists. There might be some differences in how they measure cortical thickness, transverse diameter or clinical tumour size and in how they classify (non)suspicious lymph nodes. This inter-observer variability could possibly influence the outcome of this validation study. However, inter-observer variability will be present in clinical practice, so this study would be a good representation.

Although it has been tried to use a similar study design as in the development study, small differences could not be avoided. The cut-off points used for ER-status in this study are slightly different to the cut-off points used to develop the model. The original cut-off points for patient classification are 10%, 25% and 75%, where the cut-off point used in this study are 10%, 30% and 80%. Patients with an actual ER status between 25-30% or 75-80% will have a lower predicted probability than intended, which might contributed to the underestimation of the model in some deciles. However, there is still no consensus about the relation between ER status and the risk on having ALN metastases. Some studies found no relation between ER status and ALN metastases (43, 44) and other studies found that a positive ER status is a predictor for ALN metastases. (28, 45) There is no evidence found on a linear relation between ER status and ALN metastases, which suggest that the small differences in cut-off points not really affect the outcome of this study.

This study concerns a retrospective study, whereby radiology reports from January 2011 till December 2015 were assessed. In this study, only patients with positive ultrasound findings were included. It is not known if radiologists made ultrasound images for every patient with positive ultrasound findings, and in case they did, if they imaged the most suspected lymph node present. This means that there might be patients excluded for this study, for whom there were actually ALN(s) present, but not pictured. If this comprises a specific sub-group, it could have resulted in an incomplete dataset, affecting the results of validation. This limitation could, in future research, be avoided by performing a prospective research. However, since the NCR is a large and reliable database and the patient population with imaged lymph nodes was large enough, it is believed that this study still provides representative outcomes.

Comparison with other studies

Comparable internal or external validations of different predictive models for the ALN status showed an AUC varying between 0.58 and 0.79. (27, 28, 45-49) The Memorial Sloan-Kettering Cancer Center (MSKCC) nomogram has been validated in multiple populations, including a Dutch patient population. However, on this population, the model performed moderate with an AUC of 0.67. (50) The model validated in this study already showed good discrimination in internal validation, with an AUC of 0.86, and compared to other validation studies, the AUC found in external validation shows also good discrimination. (29)

The model of Meretoja et al. shows slightly better clinical performance in external validation than found in this study. However, there were no Dutch patients included in that study, so it is not sure if that model will shows comparable performance in the Netherlands. Also, the model includes variables which cannot be assessed before primary tumour surgery, which limitation is also found in the model of Qiu et al. (29)

Implications of the findings

The model is intended to provide as additional information for both physicians, to help in decision making, and patients, to help them understand a decision made. Based on the results of this study, it is believed that the model can theoretically fulfil this goal. Although the model performs good in the Dutch situation, it does not provide the physicians with preoperative information for making further treatment decisions at this moment. Still, it might give some information for other Western European countries were histological grade and ER status is based on biopsy. The model might also be helpful for physicians in making decisions about adjuvant chemo- and/or radiotherapy for patients where axillary surgery for some reason could not be performed. Further research is needed for developing a model that can be used as preoperative diagnostic tool in the Dutch clinical practise. Also, the validation population consisted of Dutch patients only and clinical practice varies in and between countries. Therefore, validation of this model in other countries is still worthwhile.

Appendix 1: Developing a new preoperative model

This appendix was intended to include a set up of a study to develop a new predictive model for the axillary lymph node status in Dutch early stage breast cancer patients. However, due to limited time and lack of accurate knowledge about internal validation methods (bootstrapping), the development of the model is the only phase that is performed. The main section of this study forms the base for this additional part. Only additions and modifications are mentioned in this appendix.

Methods

Only additions and modifications to the method in the main section are mentioned in this method.

Setting and subjects

In this study, 395 patients diagnosed in Medisch Spectrum Twente and Isala were included.

Data collection

The longitudinal diameter was collected for all patients in addition to the variables mentioned in the main section. The transverse/longitudinal axis ratio is calculated by dividing the longitudinal diameter by the transverse diameter.

Statistical methods

For developing the model, a significance level of 15% was maintained. Univariate analyses was performed using logistic regression analyses with ALN status as dependent variable and all possible predictors separately as independent variable. All significant predictors were included in a multivariate analyses, using the backward selection method. The performance of the model was assessed using the AUC and the HL-test.

Results

Patient-, tumour- and ALN characteristics of the development population and the p values resulted from univariate analyses are shown in *Table 1*.

Table 1 Patient-, tumour- and ALN characteristics; univariate analysis (logistic regression)

Variable	**Population No. (%)** ***n=395***	**p-value**
Age at diagnosis – Median (IQR)	62 (52, 70)	0.01
Menopausal status		0.71
Premenopausal (1)	5 (13.92%)	
Postmenopausal (0)	294 (74.43%)	
Perimenopausal (2)	24 (22%)	
Unknown	22 (5.57%)	
Clinical tumour size (mm) – Median (IQR)	1.5 (1.1, 2.2)	<0.001
Tumour location		0.16
UOQ (1)	164 (41.52%)	
LOQ (2)	33 (8.35%)	

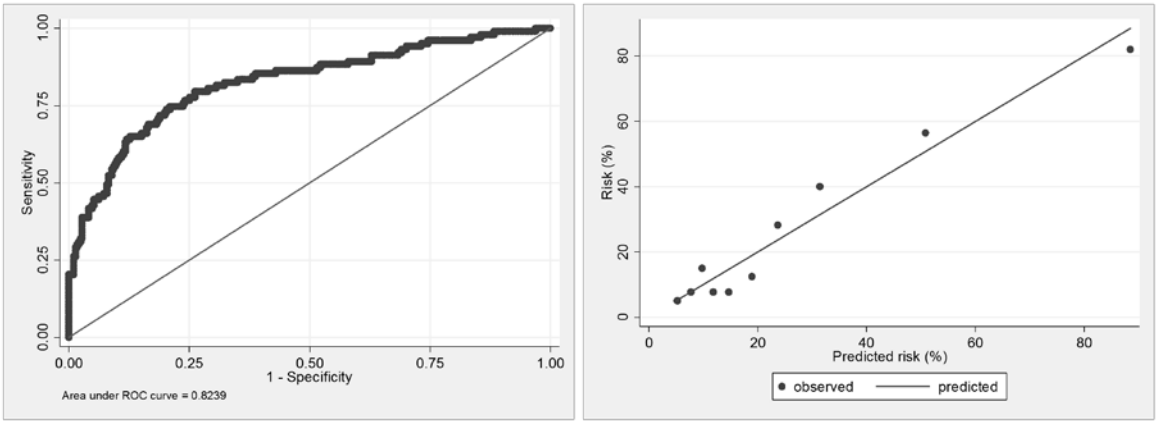
UIQ (4)	51 (12.91%)	
LIQ (3)	26 (6.58%)	
Central (5)	37 (9.37%)	
Overlapping lesions (6)	83 (21.01%)	
Unknown	1 (0.25%)	
Transverse diameter (mm) – Median (IQR)	12.9 (10, 16.5)	0.03
Longitudinal diameter (mm) – Median (IQR)	5.7 (4.5, 7.4)	<0.001
Diameter ratio – Median (IQR)	0.46 (0.37, 0.57)	0.001
Cortical thickness (mm) – Median (IQR)	1.8 (1.3, 2.7)	<0.001
Absence of hilum		<0.001
Yes	27 (6.84%)	
No	368 (93.16%)	
ALN metastases		
Yes	104 (26.33%)	
No	291 (73.67%)	

Note: UOQ: upper outer quadrant; LOQ: lower outer quadrant; UIQ: upper inner quadrant; LIQ: lower inner quadrant; IQR: interquartile range; ER: oestrogen receptor; PR: progesterone receptor; HER-2: human epidermal growth factor receptor 2; ALN: axillary lymph node.

After univariate analyses, age at diagnosis, clinical tumour size, transverse diameter, longitudinal diameter, diameter ratio, cortical thickness and absence of hilum were included in multivariate analysis. At a significance level of 15%, age, tumour size and cortical thickness remained in the model. The model is as follows:

$$ln\left(\frac{p}{1-p}\right) = -0.02 * age + 0.61 * clinical\ tumour\ size + 0.61 * cortical\ thickness - 2.30$$

The model showed an AUC of 0.82, indicating good clinical accuracy. The HL-test showed an p-value of 0.12, indicating a moderate fit for the population.



The next step

Now that a model is developed, it is important that is will be internally and externally validated. The model can be internally validated by using the bootstrap method. External validation can be performed on all patients in the dataset from the main section, who are not included in the development population. External validation can also be performed on the Chinese population, used for developing the model of Qiu et al.

5 References

1. Banerjee M, George J, Song EY, Roy A, Hryniuk W. Tree-based model for breast cancer prognostication. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2004;22(13):2567-75.

2. Soerjomataram I, Louwman MW, Ribot JG, Roukema JA, Coebergh JW. An overview of prognostic factors for long-term survivors of breast cancer. Breast cancer research and treatment. 2008;107(3):309-30.

3. Pamilo M, Soiva M, Lavast EM. Real-time ultrasound, axillary mammography, and clinical examination in the detection of axillary lymph node metastases in breast cancer patients. Journal of ultrasound in medicine : official journal of the American Institute of Ultrasound in Medicine. 1989;8(3):115-20.

4. Alvarez S, Anorbe E, Alcorta P, Lopez F, Alonso I, Cortes J. Role of sonography in the diagnosis of axillary lymph node metastases in breast cancer: a systematic review. AJR American journal of roentgenology. 2006;186(5):1342-8.

5. Cho N, Moon WK, Han W, Park IA, Cho J, Noh DY. Preoperative sonographic classification of axillary lymph nodes in patients with breast cancer: node-to-node correlation with surgical histology and sentinel node biopsy results. AJR American journal of roentgenology. 2009;193(6):1731-7.

6. Luparia A, Campanino P, Cotti R, Lucarelli D, Durando M, Mariscotti G, et al. Role of axillary ultrasound in the preoperative diagnosis of lymph node metastases in patients affected by breast carcinoma. La Radiologia medica. 2010;115(2):225-37.

7. Mainiero MB, Cinelli CM, Koelliker SL, Graves TA, Chung MA. Axillary ultrasound and fine-needle aspiration in the preoperative evaluation of the breast cancer patient: an algorithm based on tumor size and lymph node appearance. AJR American journal of roentgenology. 2010;195(5):1261-7.

8. Rahbar H, Partridge SC, Javid SH, Lehman CD. Imaging axillary lymph nodes in patients with newly diagnosed breast cancer. Current problems in diagnostic radiology. 2012;41(5):149-58.

9. NABON. Mammacarcinoom. Landelijke richtlijn, versie: 2.0. 2012.

10. Abe H, Schmidt RA, Kulkarni K, Sennett CA, Mueller JS, Newstead GM. Axillary lymph nodes suspicious for breast cancer metastasis: sampling with US-guided 14-gauge core-needle biopsy--clinical experience in 100 patients. Radiology. 2009;250(1):41-9.

11. Moore A, Hester M, Nam MW, Brill YM, McGrath P, Wright H, et al. Distinct lymph nodal sonographic characteristics in breast cancer patients at high risk for axillary metastases correlate with the final axillary stage. The British journal of radiology. 2008;81(968):630-6.

12. Deurloo EE, Tanis PJ, Gilhuijs KG, Muller SH, Kroger R, Peterse JL, et al. Reduction in the number of sentinel lymph node procedures by preoperative ultrasonography of the axilla in breast cancer. European journal of cancer. 2003;39(8):1068-73.

13. Lyman GH, Giuliano AE, Somerfield MR, Benson AB, 3rd, Bodurka DC, Burstein HJ, et al. American Society of Clinical Oncology guideline recommendations for sentinel lymph node biopsy in early-stage breast cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2005;23(30):7703-20.

14. Donker M, van Tienhoven G, Straver ME, Meijnen P, van de Velde CJ, Mansel RE, et al. Radiotherapy or surgery of the axilla after a positive sentinel node in breast cancer (EORTC 10981-22023 AMAROS): a randomised, multicentre, open-label, phase 3 non-inferiority trial. The Lancet Oncology. 2014;15(12):1303-10.

15. Nori J, Vanzi E, Bazzocchi M, Bufalini FN, Distante V, Branconi F, et al. Role of axillary ultrasound examination in the selection of breast cancer patients for sentinel node biopsy. American journal of surgery. 2007;193(1):16-20.

16. McLaughlin SA, Wright MJ, Morris KT, Giron GL, Sampson MR, Brockway JP, et al. Prevalence of lymphedema in women with breast cancer 5 years after sentinel lymph node biopsy or axillary dissection: objective measurements. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2008;26(32):5213-9.

17. Ashikaga T, Krag DN, Land SR, Julian TB, Anderson SJ, Brown AM, et al. Morbidity results from the NSABP B-32 trial comparing sentinel lymph node dissection versus axillary dissection. Journal of surgical oncology. 2010;102(2):111-8.

18. Langer I, Guller U, Berclaz G, Koechli OR, Schaer G, Fehr MK, et al. Morbidity of sentinel lymph node biopsy (SLN) alone versus SLN and completion axillary lymph node dissection after breast cancer surgery: a prospective Swiss multicenter study on 659 patients. Annals of surgery. 2007;245(3):452-61.

19. Rietman JS, Dijkstra PU, Geertzen JH, Baas P, De Vries J, Dolsma W, et al. Short-term morbidity of the upper limb after sentinel lymph node biopsy or axillary lymph node dissection for Stage I or II breast carcinoma. Cancer. 2003;98(4):690-6.

20. Kell MR, Burke JP, Barry M, Morrow M. Outcome of axillary staging in early breast cancer: a meta-analysis. Breast cancer research and treatment. 2010;120(2):441-7.

21. Mansel RE, Fallowfield L, Kissin M, Goyal A, Newcombe RG, Dixon JM, et al. Randomized multicenter trial of sentinel node biopsy versus standard axillary treatment in operable breast cancer: the ALMANAC Trial. Journal of the National Cancer Institute. 2006;98(9):599-609.

22. Fracheboud J, Otto SJ, van Dijck JA, Broeders MJ, Verbeek AL, de Koning HJ, et al. Decreased rates of advanced breast cancer due to mammography screening in The Netherlands. British journal of cancer. 2004;91(5):861-7.

23. Galimberti V, Cole BF, Zurrida S, Viale G, Luini A, Veronesi P, et al. Axillary dissection versus no axillary dissection in patients with sentinel-node micrometastases (IBCSG 23-01): a phase 3 randomised controlled trial. The Lancet Oncology. 2013;14(4):297-305.

24. Gentilini O, Veronesi U. Abandoning sentinel lymph node biopsy in early breast cancer? A new trial in progress at the European Institute of Oncology of Milan (SOUND: Sentinel node vs Observation after axillary UltraSouND). Breast. 2012;21(5):678-81.

25. Perhavec A, Perme MP, Hocevar M, Besic N, Zgajnar J. Ljubljana nomograms for predicting the likelihood of non-sentinel lymph node metastases in breast cancer patients with a positive sentinel lymph node. Breast cancer research and treatment. 2010;119(2):357-66.

26. Meretoja TJ, Strien L, Heikkila PS, Leidenius MH. A simple nomogram to evaluate the risk of nonsentinel node metastases in breast cancer patients with minimal sentinel node involvement. Annals of surgical oncology. 2012;19(2):567-76.

27. Meretoja TJ, Heikkila PS, Mansfield AS, Cserni G, Ambrozay E, Boross G, et al. A predictive tool to estimate the risk of axillary metastases in breast cancer patients with negative axillary ultrasound. Annals of surgical oncology. 2014;21(7):2229-36.

28. Xie F, Yang H, Wang S, Zhou B, Tong F, Yang D, et al. A logistic regression model for predicting axillary lymph node metastases in early breast carcinoma patients. Sensors. 2012;12(7):9936-50.

29. Qiu SQ, Zeng HC, Zhang F, Chen C, Huang WH, Pleijhuis RG, et al. A nomogram to predict the probability of axillary lymph node metastasis in early breast cancer patients with positive axillary ultrasound. Scientific reports. 2016;6:21196.

30. Integraal Kankercentrum Nederland, IKNL. Over cijfers n.d. [2016-03-01].

31. A. Fritz CP, A. Jack, K. Shanmugaratnum, L.H. Sobin, M.D. Parkin. International classification of diseases for oncology (ICD-O)(3rd ed.) Geneva, Switzerland: World Health Organisation; 2000.

32. Wittekind C, Asamura, H., Sobin, L.H. Breast Tumours (ICD-0 C50). TNM Atlas Illustrated guide to the TNM classification of Malignant Tumours. 6: UICC and John Wiley & Sons; 2014. p. 459-93.

33. Twisk JWR. Inleiding in de toegepaste biostatistiek. Amsterdam: Reed Business Education; 2014.

34. (CCMO) CCMO. Niet-WMO-onderzoek: Central Commissie Mensgebonden Onderzoek; [cited 2016]. Available from: http://www.ccmo.nl/nl/niet-wmo-onderzoek.

35. Leong SP, Shen ZZ, Liu TJ, Agarwal G, Tajima T, Paik NS, et al. Is breast cancer the same disease in Asian and Western countries? World journal of surgery. 2010;34(10):2308-24.

36. Wang B, He M, Wang L, Engelgau MM, Zhao W, Wang L. Breast cancer screening among adult women in China, 2010. Preventing chronic disease. 2013;10:E183.

37. Wang LW, Yang GF, Chen JM, Yang F, Yuan JP, Sun SR, et al. A clinical database of breast cancer patients reveals distinctive clinico-pathological characteristics: a study from central China. Asian Pacific journal of cancer prevention : APJCP. 2014;15(4):1621-6.

38. Devolli-Disha E, Manxhuka-Kerliu S, Ymeri H, Kutllovci A. Comparative accuracy of mammography and ultrasound in women with breast symptoms according to age and breast density. Bosnian journal of basic medical sciences / Udruzenje basicnih mediciniskih znanosti = Association of Basic Medical Sciences. 2009;9(2):131-6.

39. Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. Critical care medicine. 2007;35(9):2052-6.

40. Pesek S, Ashikaga T, Krag LE, Krag D. The false-negative rate of sentinel node biopsy in patients with breast cancer: a meta-analysis. World journal of surgery. 2012;36(9):2239-51.

41. Goyal A, Newcombe RG, Chhabra A, Mansel RE, Group AT. Factors affecting failed localisation and false-negative rates of sentinel node biopsy in breast cancer--results of the ALMANAC validation phase. Breast cancer research and treatment. 2006;99(2):203-8.

42. Krag DN, Anderson SJ, Julian TB, Brown AM, Harlow SP, Costantino JP, et al. Sentinel-lymph-node resection compared with conventional axillary-lymph-node dissection in clinically node-negative patients with breast cancer: overall survival findings from the NSABP B-32 randomised phase 3 trial. The Lancet Oncology. 2010;11(10):927-33.

43. Aitken E, Osman M. Factors affecting nodal status in invasive breast cancer: a retrospective analysis of 623 patients. The breast journal. 2010;16(3):271-8.

44. Fehm T, Maul H, Gebauer S, Scharf A, Baier P, Sohn C, et al. Prediction of axillary lymph node status of breast cancer patients by tumorbiological factors of the primary tumor. Strahlentherapie und Onkologie : Organ der Deutschen Rontgengesellschaft [et al]. 2005;181(9):580-6.

45. Qiu PF, Liu JJ, Wang YS, Yang GR, Liu YB, Sun X, et al. Risk factors for sentinel lymph node metastasis and validation study of the MSKCC nomogram in breast cancer patients. Japanese journal of clinical oncology. 2012;42(11):1002-7.

46. Klar M, Jochmann A, Foeldi M, Stumpf M, Gitsch G, Stickeler E, et al. The MSKCC nomogram for prediction the likelihood of non-sentinel node involvement in a German breast cancer population. Breast cancer research and treatment. 2008;112(3):523-31.

47. Chen JY, Chen JJ, Yang BL, Liu ZB, Huang XY, Liu GY, et al. Predicting sentinel lymph node metastasis in a Chinese breast cancer population: assessment of an existing nomogram and a new predictive nomogram. Breast cancer research and treatment. 2012;135(3):839-48.

48. Bevilacqua JL, Kattan MW, Fey JV, Cody HS, 3rd, Borgen PI, Van Zee KJ. Doctor, what are my chances of having a positive sentinel node? A validated nomogram for risk estimation. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2007;25(24):3670-9.

49. Reyal F, Rouzier R, Depont-Hazelzet B, Bollet MA, Pierga JY, Alran S, et al. The molecular subtype classification is a determinant of sentinel node positivity in early breast carcinoma. PloS one. 2011;6(5):e20297.

50. van la Parra RF, Francissen CM, Peer PG, Ernst MF, de Roos WK, Van Zee KJ, et al. Assessment of the Memorial Sloan-Kettering Cancer Center nomogram to predict sentinel lymph node metastases in a Dutch breast cancer population. European journal of cancer. 2013;49(3):564-71.