Thieme

# Predicting Triple-Negative Breast Cancer Subtype Using Multiple Single Nucleotide Polymorphisms for Breast Cancer Risk and Several Variable Selection Methods

## Die Vorhersage von triple-negativem Brustkrebs mithilfe von brustkrebsassoziierten Einzelnukleotid-Polymorphismen und verschiedenen Variablenselektionsmethoden

Authors

Lothar Häberle[1,2], Alexander Hein[1], Matthias Rübner[1], Michael Schneider[1], Arif B. Ekici[3], Paul Gass[1], Arndt Hartmann[4], Rüdiger Schulz-Wendtland[5], Matthias W. Beckmann[1], Wing-Yee Lo[6,7], Werner Schroth[6,7], Hiltrud Brauch[6,7,8], Peter A. Fasching[1*], Marius Wunderle[1*]

Affiliations

1 Department of Gynecology and Obstetrics, Erlangen University Hospital, University Breast Center for Franconia, Comprehensive Cancer Center Erlangen-EMN, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany
2 Biostatistics Unit, Department of Gynecology and Obstetrics, Erlangen University Hospital, Erlangen, Germany
3 Institute of Human Genetics, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany
4 Institute of Pathology, Erlangen University Hospital, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany
5 Institute of Diagnostic Radiology, Erlangen University Hospital, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany
6 Dr. Margarete Fischer-Bosch Institute of Clinical Pharmacology, Stuttgart, Germany
7 University of Tübingen, Tübingen, Germany
8 German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany

Key words

breast cancer, SNPs, triple-negative, subtype prediction, prediction model, variable selection

Schlüsselwörter

Brustkrebs, SNPs, triple-negativ, Subtypvorhersage, Prädiktionsmodell, Variablenselektion

received 7.4.2017
revised 15.5.2017
accepted 16.5.2017

Bibliography

DOI https://doi.org/10.1055/s-0043-111602
Geburtsh Frauenheilk 2017; 77: 667–678 © Georg Thieme Verlag KG Stuttgart · New York | ISSN 0016-5751

Correspondence

Lothar Häberle, PhD
Department of Gynecology and Obstetrics, Erlangen University Hospital, University Breast Center for Franconia, Comprehensive Cancer Center Erlangen-EMN, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)
Universitätsstraße 21–23, 91054 Erlangen, Germany
lothar.haeberle@uk-erlangen.de

## ABSTRACT

**Introduction** Studies of triple-negative breast cancer have recently been extending the inclusion criteria and incorporating additional molecular markers into the selection criteria, opening up scope for targeted therapies. The screening phases required for studies of this type are often prolonged, since the process of determining the molecular subtype and carrying out additional biomarker assessment is time-consuming. Parameters such as germline genotypes capable of predicting the molecular subtype before it becomes available from pathology might be helpful for treatment planning and optimizing the timing and cost of screening phases. This appears to be feasible, as rapid and low-cost genotyping methods are becoming increasingly available. The aim of this study was to identify single nucleotide polymorphisms (SNPs) for breast cancer risk capable of predicting triple negativity, in addition to clinical predictors, in breast cancer patients.

**Methods** This cross-sectional observational study included 1271 women with invasive breast cancer who were treated at a university hospital. A total of 76 validated breast cancer risk SNPs were successfully genotyped. Univariate associations between each SNP and triple negativity were explored using logistic regression analyses. Several variable selection and regression techniques were applied to identify a set of SNPs that together improve the prediction of triple negativity in addi-

* Shared last authorship.

tion to the clinical predictors of age at diagnosis and body mass index (BMI). The most accurate prediction method was determined by cross-validation.

**Results** The SNP rs10069690 *(TERT, CLPTM1L)* was the only significant SNP (corrected p = 0.02) after correction of p values for multiple testing in the univariate analyses. This SNP and three additional SNPs from the genes *RAD51B, CCND1*, and *FGFR2* were selected for prediction of triple negativity. The addition of these SNPs to clinical predictors increased the cross-validated area under the curve (AUC) from 0.618 to 0.625. Age at diagnosis was the strongest predictor, stronger than any genetic characteristics.

**Conclusion** Prediction of triple-negative breast cancer can be improved if SNPs associated with breast cancer risk are added to a prediction rule based on age at diagnosis and BMI. This finding could be used for prescreening purposes in complex molecular therapy studies for triple-negative breast cancer.

## ZUSAMMENFASSUNG

**Einleitung** Studien bei triple-negativem Brustkrebs haben die Einschlusskriterien durch die Aufnahme zusätzlicher molekularer Marker erweitert. Im Rahmen des Screenings für diese Therapiestudien wird sowohl für die Bestimmung des molekularen Subtyps als als auch für zusätzliche Biomarker-Untersuchungen ein längerer Zeitraum beansprucht, was die Behandlung verzögert. Keimbahn-Genotypen könnten bei der Vorhersage des molekularen Subtyps helfen, zumal schnelle und günstige Genotypisierungsmethoden zunehmend zur Verfügung stehen. Ziel dieser Studie war es deswegen, zu prüfen, ob Einzelnukleotid-Polymorphismen (SNPs) der Keimbahn dabei helfen können, Brustkrebspatientinnen mit triple-negativem Mammakarzinom zu identifizieren.

**Methoden** In dieser Querschnittsstudie wurden 1271 Patientinnen mit invasivem Mammakarzinom eingeschlossen. Insgesamt wurden 76 validierte Brustkrebsrisiko-SNPs erfolgreich genotypisiert. Univariate Assoziationen zwischen jedem SNP und Triple-Negativität wurden mittels logistischer Regression geprüft. Verschiedene Variablenselektions- und Regressionsmethoden wurden angewandt, um eine Gruppe von SNPs zu identifizieren, die zusammen mit den klinischen Prädiktoren Alter bei Diagnose und BMI die Prädiktion der Triple-Negativität verbessern. Mittels Kreuzvalidierung wurde die Methode mit der höchsten Genauigkeit bestimmt.

**Ergebnisse** Der SNP rs10069690 *(TERT, CLPTM1L)* war der einzige einzelne SNP, der nach p-Wert-Korrektur für multiples Testen signifikant mit Triple-Negativität assoziiert war (p = 0,02). Dieser SNP und 3 weitere in den Genen *RAD51B, CCND1* und *FGFR2* wurden ausgewählt, um gemeinsam in einem Prädiktionsmodell Triple-Negativität vorherzusagen. Die Hinzunahme dieser 4 SNPs erhöhte die kreuzvalidierte AUC von 0,618 auf 0,625. Alter bei Diagnose war bei Weitem der stärkste Prädiktor.

**Schlussfolgerung** Die Vorhersage von triple-negativem Mammakarzinom kann verbessert werden, wenn sie nicht nur auf den klinischen Prädiktoren Alter bei Diagnose und BMI basiert, sondern auch auf Brustkrebsrisiko-SNPs. Das Prädiktionsmodell könnte bei der Rekrutierung von Patientinnen für aufwendige molekulare Therapiestudien eingesetzt werden.

## Introduction

Knowledge about targeted therapies for breast cancer has improved immensely over the last two decades. These therapies have mainly been developed for – although they are not restricted to – intrinsic molecular subtypes: triple-negative breast cancer (TNBC), hormone receptor-positive breast cancer, and HER2-positive breast cancer.

As TNBC lacks hormone receptors and HER2 receptors, treatment for triple-negative breast cancer is primarily restricted to conventional chemotherapy. At the molecular level, however, TNBC is a heterogeneous disease that has different histological and molecular features. Recently, studies of TNBC have been extending the inclusion criteria and now include additional molecular markers in the selection criteria, opening up scope for targeted therapies in this subtype of breast cancer.

One example is a study of the targeted antibody–drug conjugate glembatumumab vedotin [1]. The study includes not only a requirement for the tumor to be triple-negative, but also for it to show overexpression of glycoprotein nonmetastatic melanoma protein B (gpNMB). Other examples are poly-(ADP-ribose) polymerase (PARP) inhibitor studies in patients with *BRCA1/2* mutations, based on a requirement for the tumor to be triple-negative for testing of a *BRCA1/2* mutation [2, 3]. The requirement for triple negativity was later changed to HER2 negativity.

The screening phases for studies of this type are often extended, since the process of determining the molecular subtype and carrying out additional biomarker assessment is time-consuming. This can often be a challenge to the patience of both patients and physicians. Parameters capable of predicting the molecular subtype before it becomes available via pathology might be helpful for treatment planning and for optimizing the timing and cost of screening phases for clinical trials. Biomarker assessment could be carried out at an early stage in the work-up for patients with an increased likelihood of the specific molecular subtype.

Clinical and epidemiological risk factors, such as reproductive factors and body mass index (BMI), are associated with the molecular subtype of the tumor. They appear to have an effect on the risk of developing hormone receptor-positive tumors [4–6]. In a case–case analysis, our group previously reported that age and BMI are the most important parameters associated with molecular subtypes [7]. High mammographic density was also associated with hormone receptor-negative tumors [8, 9].

Since rapid and low-cost genotyping is becoming increasingly widely available [10], single nucleotide polymorphisms (SNPs) might be useful as predictors for molecular subtypes. Genetic fac-

tors have been shown to increase the risk for specific breast cancer subtypes. For example, it is known that patients with *BRCA1* mutations are mainly diagnosed with triple-negative breast cancer, and mutation rates in this population are over 10 % [11]. In addition, approximately 100 validated SNPs for breast cancer risk are known [12–14]. Some of these SNPs have been specifically linked to a risk for hormone receptor-positive, hormone receptor-negative, or triple-negative breast cancer [15–21].

It was hypothesized that a combination of multiple breast cancer risk SNPs in addition to clinical predictors of molecular subtypes may improve the prediction of molecular subtypes. Specifically, predicting TNBC – a breast cancer subtype in which the patients affected have many unmet medical needs – would be helpful. The aim of this study was therefore to identify breast cancer risk SNPs capable of predicting TNBC in addition to clinical predictors in women with invasive breast cancer. The prediction performance of various methods of selecting SNPs was compared.

# Methods

## Patients

The patients selected for this retrospectively designed cross-sectional observational study are included in the Bavarian Breast Cancer Cases and Cohorts (BBCC) study. The BBCC has been ongoing since 2002 and includes consecutively recruited patients with invasive breast cancer at the University Breast Center for Franconia. The study was designed to identify and validate genetic and non-genetic risk factors, and it has been involved in several validation studies for SNPs [13, 14, 19–33]. For the present study, all women who were recruited into the BBCC from 2002 to 2010 were selected. Among them, patients were excluded for the following reasons: no participation in any genetic BBCC research projects; insufficient remaining DNA available due to participation in previous research projects; and no data on hormone receptor status or HER2 status available from the central pathology department at the breast cancer center. After SNPs had been selected for analysis (see below), patients with incomplete genetic information were also excluded. All of the patients provided written informed consent, and the Ethics Committee of the Medical Faculty at Friedrich Alexander University of Erlangen–Nuremberg approved the study.

## Data collection

All treatment-related patient data and tumor characteristics were documented as part of the certification processes required by the German Cancer Society *(Deutsche Krebsgesellschaft)* and by the German Society for Breast Diseases *(Deutsche Gesellschaft für Senologie)* [34]. The data are recorded prospectively in a database and audited annually as part of the breast cancer center certification process. Epidemiological data and risk factors for breast cancer were obtained using a structured questionnaire, which was completed by the patients and reviewed together with trained study personnel and supplemented if necessary.

## SNP selection

A total of 102 SNPs were selected for genotyping. Of these, 98 are validated breast cancer risk SNPs. Most of these breast cancer risk SNPs have been confirmed in large international validation studies, mainly by the Breast Cancer Association Consortium (BCAC). The BCAC initially published a validation of a few SNPs [35] and then, after increasing the sample size and analyzing more SNPs, published a series of papers as a result of the Collaborative Oncological Gene–environment Study (COGS; www.cogseu.org and www.nature.com/icogs/) [13, 14]. Four SNPs that were shown to have an influence on the prognosis in breast cancer patients were also selected [36–39]. A complete list of the SNPs, including references, is provided in Supplementary Table **S1** [13, 14, 18–20, 25, 29, 30, 32, 33, 35, 37–39, 67, 79–89].

## DNA extraction, genotyping, and quality control

Whole-blood samples were collected in citrate-phosphate-dextrose-adenine (CPDA) tubes (Sarstedt AG, Nümbrecht, Germany) from patients who had consented to participate in the biomarker substudy. Germline DNA was extracted using the automated magnetic bead-based chemagic MSM I technique (PerkinElmer chemagen, Baesweiler, Germany) in accordance with the manufacturer's instructions. Genotyping was done at the Dr. Margarete Fischer-Bosch Institute of Clinical Pharmacology, using MassARRAY iPLEX Gold (Sequenom, San Diego, California, USA). SNPs were excluded if MALDI spectra were unreliable, based on raw genotype data. Exact tests for Hardy–Weinberg equilibrium (HWE) were performed and SNPs with an unexpectedly small p value, assessed using a quantile–quantile plot, were also excluded.

## Pathology assessment

All of the histopathological information used in the analysis was directly documented from the original pathology reports, which were reviewed by two investigators. Estrogen receptor status, progesterone receptor status, and HER2 status were assessed as follows. Monoclonal mouse antibodies against estrogen receptor-alpha (clone 1D5; 1 : 200 dilution, DAKO, Denmark) and monoclonal mouse antibody against the progesterone receptor (clone pgR636, 1 : 200 dilution, DAKO, Denmark) were used to stain the pretreatment core biopsies. The percentage of positively stained cells was included in the pathology reports. The tumors were considered to be positive for the estrogen and progesterone receptors if 10 % or more of the cells showed positive staining, in accordance with recommendations applying at the time when the study was conducted [40–43]. A polyclonal antibody against HER2 (1 : 200 dilution, DAKO, Denmark) was used, and HER2 status was stated in the pathology reports as negative, 0, 1+, 2+, or 3+ in accordance with the guidelines published by Sauter et al. [44]. Tumors with a score of 0 or 1+ were regarded as HER2-negative, and those with a score of 3+ were regarded as HER2-positive. Tumors with 2+ staining were tested for gene copy numbers of HER2 by chromogene in-situ hybridization. Using a kit with two probes of different colors (ZytoDot, 2C SPEC HER2/CEN17, ZytoVision Ltd., Bremerhaven, Germany), the gene copy numbers of HER2 and centromeres of the corresponding chromosome 17 were retrieved. A HER2/CEN17 ratio of ≥ 2.2 was considered as

amplification of HER2. Scoring was carried out in a standardized way by a group of dedicated pathologists in routine surgical pathology. A tumor was regarded as being triple-negative if the estrogen receptor (ER) status was negative, progesterone receptor (PR) status was negative, and HER2 status was negative. In the present study, "triple-negative" refers to one subgroup of molecular subtypes of breast cancer, although comprehensive gene expression profiling was not performed.

## Statistical methods

To investigate the predictive value of each single SNP relative to the occurrence of a TNBC in addition to clinical predictors, a multiple logistic regression model was fitted for each SNP with TNBC status (yes versus no) as the outcome, and with the specific SNP (ordinal; 0, 1, or 2 minor alleles) and the clinical predictors age at diagnosis (continuous) and BMI (continuous) as predictors [7]. Patients with missing genetic data or missing outcome data were excluded. Missing clinical predictors were imputed, as done in [45]. Continuous predictors were used as natural cubic spline functions to describe nonlinear effects [46].The degrees of freedom (between 1 and 3) of each predictor were calculated as done in [45]. The odds ratio (OR) per minor allele with confidence interval was calculated using the logistic regression model. For each SNP, a likelihood ratio test comparing the clinical-genetic logistic regression model with a clinical logistic regression model containing only the clinical predictors was performed. The p values (one per SNP) were corrected for multiple testing using the Bonferroni–Holm method.

The primary study aim was to identify a set of SNPs that together would improve the prediction of TNBCs in addition to clinical predictors (age, BMI). Identifying relevant SNPs among the relatively large number of candidate SNPs was a challenging process, which can be summed up as follows. The complete dataset was randomly divided into two parts: one training set with about two-thirds of the patients, and one validation set with about one-third of the patients. Various SNP selection methods and regression techniques, respectively, were applied to the training data to obtain regression models with selected SNPs and clinical predictors. The models were compared among themselves with regard to their prediction errors on validation data.

All but one of the regression techniques considered comprise a bundle of candidate models characterized by a tuning parameter λ. The optimal λ has to be determined before a specific prediction model representing the regression technique can be fitted to predict TNBC. After the degrees of freedom of the continuous clinical predictors had been determined again by using training data, the following regression techniques were applied to the training data.

*Univariate selection.* For each SNP, a logistic regression model with the clinical predictors and the specific SNP was compared with a logistic regression model with clinical predictors alone, using a likelihood ratio test. The SNPs were ordered according to increasing p values for these likelihood ratio tests. The λ top-ranked SNPs were selected and included in a logistic regression model that also contained the clinical predictors. Here λ, ranging from 0 to 30, is a tuning parameter representing the number of selected SNPs [47]. When a specific model was applied to the validation data afterwards, generalized shrinkage after coefficient estimation toward the clinical regression model was used to improve predictions [48].The shrinkage factor was obtained from the maximal genetic model with 30 SNPs.

*Stepwise selection* as described in [49]. All of the SNPs were ordered as above. The top 30 ranked SNPs were preselected, in order to keep the number of SNPs to be analyzed easy to handle. One hundred bootstrap samples of the same size as the original dataset were drawn with replacement. On each bootstrap sample, a logistic regression model with the clinical predictors and the preselected SNPs was set up. A backward stepwise variable selection procedure that kept all the clinical predictors was carried out to obtain the best model in accordance with the Akaike information criterion. The retained variables from each bootstrap sample were recorded, and a final variable selection was made. The most frequently selected SNPs (> 70 %) and – to address correlation among SNPs – representatives of highly frequent SNP pairs (> 90 %) were chosen. Again, generalized shrinkage was incorporated when the final model was applied.

The *least absolute shrinkage operator (lasso)* is a regression technique in which the regression coefficients are shrunk towards zero during estimation [50]. The amount of shrinkage is controlled by a tuning parameter λ. Depending on the value of λ, a number of coefficients reach exactly zero, which means that lasso also leads to variable selection. In the present study, a regression model was set up with the clinical predictors and all SNPs. The coefficients of the SNPs, but not the coefficients of the clinical predictors, were shrunk by variation of λ. A regression model with maximal shrinkage that has all coefficients of the SNPs equal to 0 corresponds to the clinical logistic regression model. In contrast to the usual regression models, lasso can deal with large numbers of predictors.

Component-wise gradient *boosting* fits a regression model iteratively [51, 52]. It starts with an empty model without any predictors. In each iteration, the best-performing predictor is added to the model with a small step size, or its coefficient is updated if it was included before. More relevant predictors are included earlier than less relevant ones. The number of boosting iterations, λ, is a tuning parameter that controls both the variable selection properties of the algorithm and the implied shrinkage of the coefficients. The incorporation of clinical predictors is less straightforward than for lasso. A logistic regression model with clinical predictors is fitted. This fit is taken as the offset for the boosting procedure described above with SNPs as predictors [53].

The optimal λ for each method was found by 10-fold cross-validation on the training dataset. For a given value of λ, the prediction model was estimated on nine folds and then applied on the tenth fold. The mean squared error (MSE) was taken as the evaluation measure. The MSE is a summary measure of the differences between the observed TNBC status (either 0 for "no" or 1 for "yes") of patients in the tenth fold, which was not used for model building, and the expected probability obtained from the model (between 0 and 1) for these patients having a TNBC. This procedure was done 10 times, leaving one fold out at a time, and the average MSE was calculated. The λ value with the smallest average MSE was regarded as the optimal λ. The whole training set was finally used to fit a regression model with the optimal λ.

The procedures described above resulted in four clinical-genetic regression models for predicting TNBC. In addition, two

benchmark models – a logistic null model without any predictors and a clinical logistic regression model with clinical predictors but without any SNPs – were fitted on the training data. A useful clinical model should perform better than the null model, whereas a useful prediction model with clinical and genetic predictors should perform better than the clinical model without further predictors. These six models were evaluated on the validation dataset to measure their performance in new patients. Again, the MSE was taken as a performance criterion.

To obtain further insight into the accuracy of the prediction, the performance improvement of the four clinical-genetic models in comparison with the clinical model was assessed on validation data using the continuous net reclassification improvement (NRI). Roughly speaking, the continuous NRI is the proportion of patients with TNBC or without TNBC, respectively, who are correctly given a higher or lower predicted probability of TNBC by the clinical-genetic regression model than by the clinical model, corrected by wrongly assigned lower or higher probabilities [54].

In clinical practice, a prediction model for TNBC might support treatment decision-making based on a threshold for the predicted probability of TNBC that classifies a patient as a "high-risk" patient or "low-risk" patient. The ability to distinguish between patients with and without TNBC was measured on validation data using the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC), an estimation of the probability that given two patients, one with TNBC and the other without TNBC, the prediction model will assign TNBC status to both patients correctly.

To overcome the drawbacks of only splitting the data into training and validation sets once, the dataset was divided several times into training and validation sets and the procedure was repeated as described above each time [47, 55]. More precisely, 3-fold cross-validation with 100 repetitions was done. For each regression technique for predicting TNBC, the average value of the 300 MSEs of the corresponding regression models was taken as a final evaluation criterion, and the average AUC and average NRI were used as further criteria. The regression technique with the smallest average MSE is regarded as the best method (the "winner" method) for predicting TNBC.

The best prediction method was applied to the whole dataset to obtain the final prediction model for TNBC. This was done by repeating all of the model-building steps, this time not on the training data, but on the complete dataset. That is, cubic spline functions and the tuning parameter λ were determined as described above and a corresponding regression model was fitted on the complete dataset. A TNBC prediction score on a scale from 0 to 100, representing the probability of a TNBC, was derived from the final prediction model by taking the inverse logit of the linear combination of predictor values and regression coefficients. The performance of the final model on the complete dataset in terms of discrimination and calibration was measured using the AUC and the Hosmer–Lemeshow statistic (scatterplot and $\chi^2$ test) comparing predicted and observed TNBC events, as done recently in [56]. A large p value indicates satisfactory calibration.

All of the tests were two-sided, and a p value < 0.05 was considered statistically significant. Calculations were carried out using the R system for statistical computing (version 3.0.1; R Core Team, Vienna, Austria, 2013).

▶ **Table 1** Patient characteristics.

| Characteristic | | Mean or count | SD or % |
|---|---|---|---|
| Age at diagnosis | Years | 57.2 | 11.7 |
| BMI | kg/m$^2$ | 26.2 | 4.8 |
| ER | ▪ Negative | 231 | 22.5 |
| | ▪ Positive | 796 | 77.5 |
| PR | ▪ Negative | 288 | 28.0 |
| | ▪ Positive | 739 | 72.0 |
| HER2 | ▪ Negative | 877 | 85.4 |
| | ▪ Positive | 150 | 14.6 |
| Triple-negative | ▪ No | 893 | 87.0 |
| | ▪ Yes | 134 | 13.0 |

BMI, body mass index; ER, estrogen receptor; PR, progesterone receptor

# Results

## Patients and SNPs

A total of 2234 patients were recruited into the BBCC during the specified period. A subset of 1868 patients took part in genetic BBCC research projects. Of these, sufficient DNA was available from 1743 patients. A further 472 patients with incomplete hormone receptor and HER2 status information were excluded, resulting in 1271 remaining patients. Twenty-seven out of 102 SNPs were excluded after genotype quality control: 24 SNPs because of unreliable MALDI spectra and three SNPs because of departure from HWE (Supplementary Table **S1**). Due to missing values, the following SNPs were excluded: rs1550623 (17.0% missing values out of 1271), rs3903072 (9.9%), rs2380205 (9.6%), rs17817449 (7.4%), rs2236007 (7.2%), rs3803662 (5.0%), rs9790517 (5.0%), and rs2046210 (5.0%). All patients had age information, and missing BMI values (4.2%) were imputed. The final sample size was 1027 patients, after 244 patients with incomplete genetic information had been excluded. Patient characteristics are shown in ▶ **Table 1**.

## Univariate SNP and TNBC association

The clinical predictors age at diagnosis and BMI, used as adjustment variables, fitted best as cubic spline functions with 2 and 1 degrees of freedom, respectively – i.e., age was used nonlinearly and BMI was used linearly. Twenty SNPs with the smallest p values in the univariate analyses are shown in ▶ **Table 2**. rs10069690 *(TERT, CLPTM1L)* was the only significant SNP (corrected p = 0.02) after correction of p values for multiple testing. The corrected p values for rs2981579 *(FGFR2)*, rs7726159 *(TERT)*, rs2588809 *(RAD51B)*, and rs78540526 *(CCND1)* were 0.18, 0.36, 0.81 and 0.93, respectively; the other corrected p values were 1.00.

▶ **Table 2** Univariate associations with triple-negative breast cancer (TNBC) for the 20 SNPs with the lowest p values.

| SNP | Chromosome | Nearest genes | MAF | OR (95% CI)[1] | p value[2] |
|---|---|---|---|---|---|
| rs10069690 | 5 | *TERT, CLPTM1L* | 0.249 | 1.66 (1.27, 2.18) | < 0.001 |
| rs2981579 | 10 | *FGFR2* | 0.484 | 0.66 (0.51, 0.87) | < 0.01 |
| rs7726159 | 5 | *TERT* | 0.358 | 1.46 (1.12, 1.91) | < 0.01 |
| rs2588809 | 14 | *RAD51B* | 0.174 | 0.62 (0.42, 0.92) | 0.02 |
| rs78540526 | 11 | *CCND1* | 0.104 | 0.55 (0.32, 0.92) | 0.02 |
| rs11820646 | 11 | – | 0.389 | 1.38 (1.06, 1.80) | 0.02 |
| rs2981582 | 10 | *FGFR2* | 0.451 | 0.73 (0.55, 0.95) | 0.02 |
| rs3760982 | 19 | *KCNN4, ZNF283* | 0.485 | 0.77 (0.59, 1.01) | 0.06 |
| rs2363956 | 19 | *MERIT40* | 0.488 | 0.78 (0.60, 1.01) | 0.06 |
| rs1436904 | 18 | *CHST9* | 0.383 | 1.27 (0.98, 1.65) | 0.07 |
| rs6001930 | 22 | *MKL1* | 0.127 | 0.68 (0.43, 1.06) | 0.09 |
| rs12422552 | 12 | *ATF7IP, GRIN2B* | 0.295 | 0.77 (0.57, 1.04) | 0.09 |
| rs8170 | 19 | *MERIT40* | 0.191 | 1.31 (0.97, 1.78) | 0.08 |
| rs941764 | 14 | *CCDC88C* | 0.354 | 0.79 (0.59, 1.04) | 0.09 |
| rs11075995 | 16 | *FTO* | 0.264 | 1.29 (0.96, 1.72) | 0.09 |
| rs12710696 | 2 | – | 0.357 | 1.26 (0.96, 1.66) | 0.09 |
| rs11365234 | 7 | *AKAP9* | 0.392 | 1.24 (0.96, 1.61) | 0.10 |
| rs2823093 | 21 | *NRIP1* | 0.275 | 1.26 (0.96, 1.67) | 0.10 |
| rs4666275 | 2 | *ALK* | 0.060 | 1.50 (0.92, 2.46) | 0.11 |
| rs75915166 | 11 | *CCND1* | 0.082 | 0.67 (0.40, 1.14) | 0.14 |

SNP, single nucleotide polymorphism; MAF, minor allele frequency

[1] Odds ratio (OR) per minor allele, adjusted for age and body mass index, with 95% confidence interval (CI) and corresponding p value, obtained from the multiple logistic regression model.

[2] Uncorrected p values. The corrected p values for the top five SNPs were 0.02, 0.18, 0.36, 0.81, and 0.93. All other corrected p values were 1.00.

▶ **Table 3** Prediction of triple-negative tumor[1].

| Model | MSE | Reclassification (%) | | | AUC | Selected SNPs |
|---|---|---|---|---|---|---|
| | | NRI | Correctly up | Correctly down | | |
| Null[2] | 0.1137 (0.0109) | – | – | – | 0.500 (0.000) | – |
| Clinical[3] | 0.1098 (0.0104) | – | – | – | 0.618 (0.036) | – |
| Univariate selection[4] | 0.1098 (0.0107) | 9.0 (12.2) | 29.9 (25.2) | 35.3 (29.1) | 0.620 (0.038) | 2.2 (2.9) |
| Stepwise selection[4] | 0.1108 (0.0108) | 13.8 (13.9) | 46.0 (7.5) | 60.6 (5.4) | 0.614 (0.037) | 8.1 (2.5) |
| Lasso[4] | 0.1096 (0.0103) | 12.5 (16.7) | 49.1 (19.9) | 57.1 (16.9) | 0.622 (0.039) | 9.1 (7.5) |
| Boosting[4] | 0.1095 (0.0103) | 17.3 (13.8) | 55.4 (9.4) | 53.3 (8.5) | 0.625 (0.037) | 8.2 (7.2) |

AUC, area under the curve; MSE, mean squared error; NRI, net reclassification improvement; SNP, single nucleotide polymorphism

[1] Summary statistics (mean and standard deviation) for MSE, NRI, and AUC, obtained from (logistic) regression models as well as the number of selected SNPs are shown. All measures were obtained by 3-fold cross-validation with 100 repetitions.

[2] Logistic regression model without any predictors.

[3] Logistic regression model with clinical predictors (age and body mass index), but without any genetic predictors.

[4] Regression model with clinical predictors and selected SNPs.

## Clinical-genetic TNBC prediction

Boosting turned out to be the most accurate prediction method, and had a slightly smaller cross-validated prediction error MSE than the lasso (▶ **Table 3**). Lasso and boosting performed better than the clinical prediction model without genetic predictors, whereas univariate selection performed similarly and stepwise selection performed less well. These results were confirmed by AUC statistics: Boosting was also superior with regard to distinguishing between TNBC patients and non-TNBC patients. Lasso and univari-

ate selection performed better than the clinical model, and stepwise selection less well.

Boosting correctly increased the predicted probabilities of TNBC for the majority of patients with a TNBC ("correct reclassification upwards" in ▶ **Table 3**) and correctly decreased the predicted probabilities of TNBC for the majority of patients without TNBC ("correct reclassification downwards" in ▶ **Table 3**). Lasso did these correct increases and decreases for about half of the TNBC patients and the majority of the non-TNBC patients. Univariate selection correctly increased and decreased prediction probabilities only for a minority of patients. With regard to correct reclassifications, stepwise selection performed much better than univariate selection. In total, the reclassification improvement of the boosting model was superior to all other methods ("NRI" in ▶ **Table 3**).

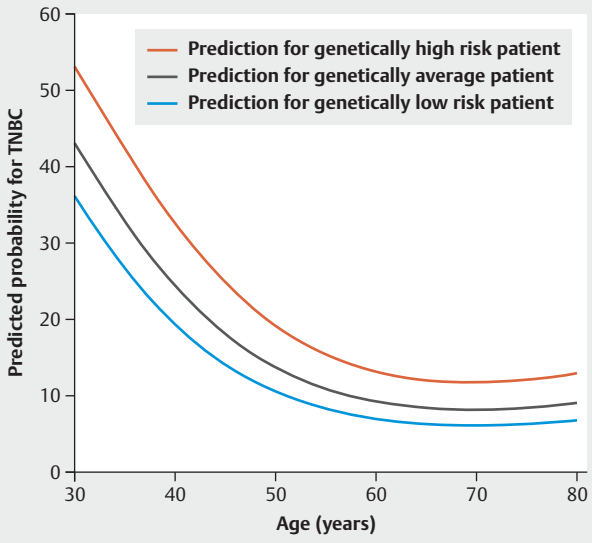
The average number of selected SNPs on the 300 training samples was similar at boosting, lasso, and stepwise selection and smaller at univariate selection. The number of SNPs varied relatively strongly at lasso and boosting and weakly at stepwise selection and univariate selection (▶ **Table 3**).

During cross-validation, univariate tests were performed on each training set and SNPs were ordered according to their p values. The most frequent SNP on top was rs10069690, ranking first 158 times (52.7%). The next most frequent SNPs on top were rs2981579 (17.7%), rs78540526 (5.7%), rs2588809 (4.3%), and rs7726159 (4.3%). In total, 24 SNPs were ranked first at least once.

A boosting prediction model, the "winner" in the method comparison, was fitted on the complete dataset. Four SNPs were selected: rs10069690 *(TERT, CLPTM1L)*, rs2981579 *(FGFR2)*, rs2588809 *(RAD51B)*, and rs78540526 *(CCND1)*. All of these belonged to the top five SNPs at the univariate analysis. Age was the strongest predictor, stronger than any genetic predictors. The predicted probability for TNBC as a continuous function of age is shown in ▶ **Fig. 1**. The likelihood of TNBC decreases with increasing age up to about 60 years and remains constantly low thereafter. All regression coefficients are shown in ▶ **Table 4**. The coefficients of the predictor age were approximated using a cubic polynomial, as cubic spline functions are difficult to use. Apart from age, positive coefficients are associated with an increased likelihood of TNBC. An ideal "genetically high-risk patient" can thus be defined as a patient with two minor rs10069690 alleles and always two common alleles at the other SNPs, while an ideal "genetically low-risk patient" is a patient with two common rs10069690 alleles and minor alleles at the other SNPs. The footnote in ▶ **Table 4** states how the predicted probability of TNBC can be calculated using the predictor values given.

The boosting model was well calibrated. The difference between actual and predicted events was quite low (▶ **Fig. 2**; p = 0.73, Hosmer–Lemeshow test). The apparent AUC – i.e., the AUC on the complete dataset – was 0.668, which is 0.043 units larger than the cross-validated AUC value. This indicates that the prediction model was slightly overfitted. For comparison, the apparent AUC of the clinical model was 0.632 – i.e., 0.014 units larger than its cross-validated value.

To demonstrate a possible future application of the final prediction model, various cut-off points for the TNBC risk between 0 and 100% were defined – e.g., 12%. Patients were classified as "low-risk" if the prediction model assigned a TNBC risk below 12%. Otherwise they were classified as "high-risk." The sensitivity (i.e., the proportion of patients classified as "high-risk" among true TNBC patients) and the specificity (i.e., the proportion of patients classified as "low-risk" among true non-TNBC patients) are presented in ▶ **Table 5**, and compared with the clinical model.

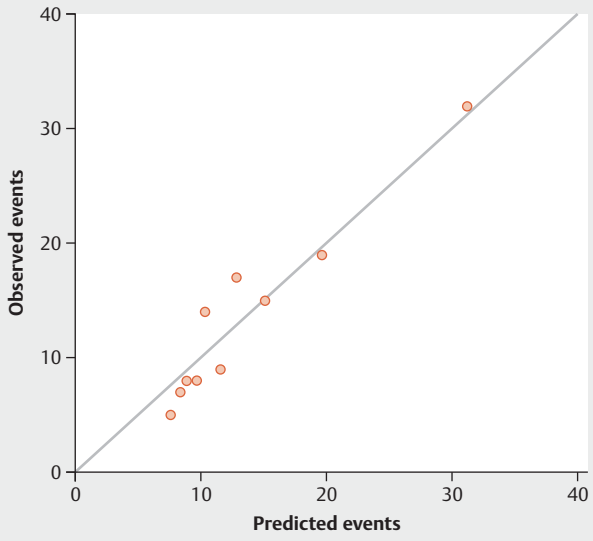



▶ **Fig. 1** The predicted probability for triple-negative breast cancer (TNBC) as a continuous function of age at diagnosis. The curves were generated using the boosting model fitted on the complete dataset. The black curve predicts the TNBC risk of a genetically "average" woman with a median body mass index. The blue and the orange curves show the predicted risk for patients with genetically maximally increased and maximally decreased risks.

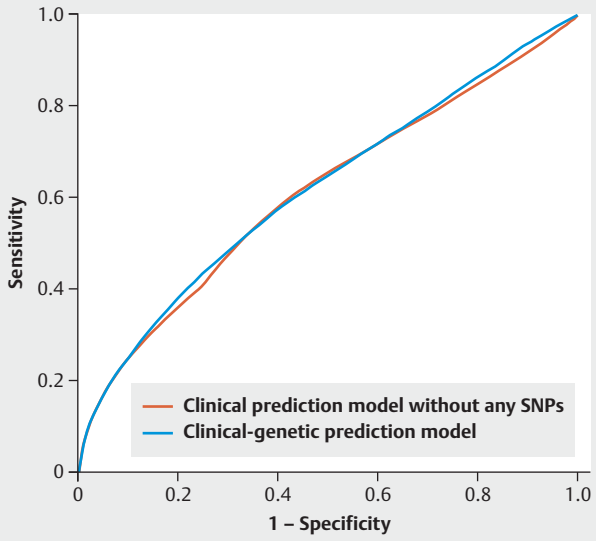
▶ **Table 4** The final clinical-genetic prediction model for triple-negative breast cancer[1].

| Predictor | Unit | Coefficient |
|---|---|---|
| Intercept | | 3.5589 |
| Age at diagnosis | Year | − 0.1624 |
| | $Year^2$ | 0.0009372 |
| | $Year^3$ | 0.000001951 |
| BMI | Per $kg/m^2$ | 0.005691 |
| rs10069690 | Minor allele | 0.19926 |
| rs2981579 | Minor allele | − 0.09108 |
| rs2588809 | Minor allele | − 0.02625 |
| rs78540526 | Minor allele | − 0.03166 |

[1] For example, the predicted probability for a 50-year-old patient with a body mass index of 26 and 1, 2, 1, 0 minor alleles of rs10069690, rs2981579, rs2588809, and rs78540526, respectively, is exp (z)/(1+exp[z]) at $z = 3.5589 + 50 \times (-0.1624) + 50^2 \times 0.0009372 + 50^3 \times 0.000001951 + 26 \times 0.005691 + 1 \times 0.19926 + 2 \times (-0.09108) + 1 \times (-0.02625) + 0 \times (-0.03166) = -1.8354$. That is, 13.8%.




▶ **Fig. 2** The observed and predicted frequencies of triple-negative breast cancer (TNBC). The patients were sorted according to the predicted probability for TNBC using the boosting prediction model and grouped into ten categories based on percentiles. The number of actually observed TNBCs ("observed events") in each category and the sum of predicted probabilities for TNBC ("predicted events") in each category are shown. Points below the gray line indicate when the model is overestimating the likelihood of TNBC; points above the gray line indicate when the model is underestimating the likelihood. A perfect prediction model would show all of the points on the gray line.




▶ **Fig. 3** Cross-validated receiver operating characteristic (ROC) curve for the clinical and clinical-genetic boosting prediction models.

The sensitivities were almost equal for cut-off points up to 12%. Thereafter, the sensitivities of the boosting model were larger. For instance, if a physician decides to screen patients with a risk of TNBC of more than 15% for biomarkers that are important for TNBC patients, without yet knowing their receptor status, then 43% of all TNBCs will be detected with the boosting model, in comparison with 38% with the clinical model. The rate of false-positive classifications would be 24%, two percentage points more than when using the clinical prediction model. The ROC curves for all possible cut-off points are shown in ▶ **Fig. 3**.

## Discussion

The study shows that prediction of TNBC can be improved if breast cancer risk SNPs are added to a prediction rule based on age at diagnosis and BMI. Age at diagnosis turned out to be the strongest predictor, stronger than any genetic influencing factors.

The final prediction model included four SNPs from the genes *RAD51B*, *TERT*, *CCND1*, and *FGFR2*. Only one of these was statistically significant in the univariate SNP and TNBC association tests, but all of them belong to the top five SNPs with the lowest p val-

▶ **Table 5** Sensitivity and specificity for the clinical prediction model and clinical-genetic boosting prediction model[1].

| Cut-off point for predicting triple-negative tumor (%)[2] | Frequency above cut-off point (%)[3] | Sensitivity | | Specificity | |
|---|---|---|---|---|---|
| | | Clinical model | Clinical-genetic model | Clinical model | Clinical-genetic model |
| 10 | 56.3 | 0.68 | 0.69 | 0.47 | 0.44 |
| 12 | 39.0 | 0.53 | 0.57 | 0.65 | 0.61 |
| 15 | 23.8 | 0.38 | 0.43 | 0.78 | 0.76 |
| 20 | 13.0 | 0.26 | 0.29 | 0.89 | 0.87 |
| 25 | 7.3 | 0.17 | 0.20 | 0.95 | 0.93 |

[1] All measurements were obtained by 3-fold cross-validation with 100 repetitions.

[2] Patients were classified into a „high-risk" group if the prediction model assigned a triple-negative tumor probability above the cut-off point. Sensitivity (between 0 and 1) is defined as the proportion of „high-risk" patients among TNBC patients. Specificity (between 0 and 1) is defined as the proportion of „low-risk" patients among non-TNBC patients.

[3] The proportion of patients classified as „high-risk" in the total study population, using the clinical-genetic prediction model.

ues. Although the selection procedure did not consider any external biological information, there might be biological reasons why these SNPs taken together improve prediction.

rs10069690 *(TERT)* has been described as being associated with estrogen receptor-negative and triple-negative breast cancer, serous ovarian cancer, breast and ovarian cancer risk in *BRCA1* mutation carriers, as well as prostate cancer – implying that there are similar pathways of pathogenesis in these different types of cancer [13, 15, 30, 33, 57]. Fine mapping analyses of this region revealed a function for telomere stability [30, 57]. rs2981579 *(FGFR2)* has been clearly described as an SNP that specifically increases the risk for hormone receptor-positive breast cancer [21, 58, 59]. Its role in hormone receptor signaling has been linked to *FOXA1*.

rs2588809 *(RAD51B)* is associated with triple-negative breast cancer [13, 15]. *RAD51B, RAD51C,* and *RAD51D* are *RAD51* paralogues that build complexes among one other [60, 61] and have a function in homologous recombination. Breast cancer in men [62], prostate cancer risk [63], and an increased risk of breast and ovarian cancer in *BRCA1* mutation carriers [64] are associated with SNPs in *RAD51B*. In vitro experiments have shown that a reduction in *RAD51B* by silencing RNA increases the chemosensitivity and reduces the efficacy of homologous recombination in breast cancer cells, with differences depending on subtype [65].

rs78540526 *(CCND1)* is located in a gene region that maps to a putative enhancer of *CCDN1*. It is clearly associated only with hormone receptor-positive breast cancer risk [25, 66, 67] and is therefore a reasonable marker for predicting hormone-receptor negativity and triple negativity. Functionally different *CCDN1* expression levels have been shown to be different with regard to haplotypes in this enhancer region [25]. This is of special interest, as *CCND1* expression and/or amplification have been under discussion as a biomarker for the efficacy of CDK4/6 inhibitors [68].

In genetic prediction studies, it can be expected that the ranking of the SNPs will differ, and the set of SNPs selected for prediction will also differ, if the experiment is repeated on a different group of patients with the same clinical characteristics. This also holds if analyses are performed on subsets of patients within one study [69]. In the present study, for instance, the top-ranked SNP in the complete dataset was *not* ranked top in about 50 % of all subsamples, and the sets of selected SNPs varied strongly. Correlations among SNPs, and SNPs with weak individual associations with the outcome but stronger power as a group, may encourage fluctuation in SNP selection. To obtain stable, reliable results that are independent of a randomly chosen patient subset, all decisions (e.g., the choice of tuning parameter for model specification and comparison of model performances) were based on repeated sampling.

Double cross-validation was carried out, with an inner loop to specify the prediction model and an outer loop to compute model performance measures, in order to ensure that all model-building steps were performed completely independently of the validation step [55, 70]. That is, all reported measures were based on data that were not used for model building. Otherwise, the measures would have been overoptimistic. Schild et al. [71] provide an example of double (cross-)validation being applied in a gynecological study.

The SNP selection process was carried out following a prespecified plan. Univariate selection is a simple method that does not take correlations among SNPs into account. It is known to perform less well in general than more sophisticated methods such as lasso and boosting [47], a result that was confirmed in this study. Lasso and boosting performed similarly, although the model fitting was rather different. However, the two methods share the common feature that variable selection is a continuous process that leads to "weakly" selected SNPs in addition to strong predictors. The result in the present study showing that boosting had a slightly better prediction accuracy is consistent with a recently published methodological study comparing boosting and lasso on simulated datasets [72]. Bootstrap-based stepwise selection, a method that our group has previously applied successfully to nongenetic data (e.g., [45, 73, 74]), performed less well than lasso and boosting. This might be because the parameters for variable selection were kept firm, in contrast to the varying tuning parameters of the other methods. Since repetitive stepwise selection is itself relatively elaborate, it would have been computationally demanding if the number of selection processes had been further increased.

The added value provided by breast cancer SNPs to a clinical prediction model was assessed using the overall performance measures MSE and AUC. The advantage of such overall measures is that prediction models can easily be compared. The disadvantage is that they may be insensitive to detecting improvements in the model performance when new predictors are added to a model that has already included important predictors [75, 76]. For example, in [77], the addition of a significant biomarker score to a set of standard risk factors increased the AUC only from 0.76 to 0.77, an increase that is similar to that in the present study. Because of this, different methods of quantifying the improvement such as the NRI have been developed [78].

In the future, germline genetic testing of SNPs from blood could be carried out in clinical routine work on the same day and at reasonable cost [10], particularly if only a few SNPs are involved that can therefore be genotyped using polymerase chain reaction. This would mean that the data would be available long before the processing of tissue, which has to be embedded, cut, and examined by a pathologist along with the relevant molecular tests. Using this genetic method of information screening for specific TNBC studies with elaborate biomarker assessment could be initiated at an early time point for patients with an increased likelihood of TNBC, particularly when biomarker assessment for all patients would be too expensive and waiting for results to come from pathology would delay biomarker assessment and the patient's entry into a study.

The present study also aimed to demonstrate ways of managing the abundance of data available in the era of "big data" and easy access to a variety of data, in order to make it feasible to use large data volumes for clinical purposes. It can be anticipated that it will also become possible to add the analysis of other markers, such as circulating tumor DNA, in order to increase the accuracy of molecular subgroup prediction. However, that will be a task for future research.

This study has some limitations. First of all, it needs to be borne in mind that the study was conducted in a population consisting only of breast cancer patients. It did not serve to identify SNPs ca-

pable of predicting the risk for triple-negative breast cancer in healthy women – e.g., using a case–control study design. As the study was intended to differentiate between triple-negative patients and non–triple-negative ones, it might have been more useful to examine SNPs differentiating between molecular subtypes rather than SNPs for breast cancer risk. Another limitation is the small sample size. With just over 1000 patients, the sample size was rather low and the findings will require validation in other independent populations.

In conclusion, the ability to predict triple-negative tumors can be improved for breast cancer patients if breast cancer risk SNPs are added to a prediction rule based on age at diagnosis and BMI. This finding could be used for prescreening purposes in complicated molecular therapy studies for triple-negative breast cancer. The advanced statistical procedures used in this study follow a prespecified, systematic plan and are described with sufficient generality to be easily adaptable for other research purposes.

## Acknowledgement

The authors are grateful to Michael Robertson for professional medical editing services.

### Conflict of Interest

The authors declare that they have no conflict of interest.

### References

[1] Celldex Therapeutics. Study of Glembatumumab Vedotin (CDX-011) in patients with metastatic, gpNMB over-expressing, triple negative breast cancer (METRIC). 2017. Online: https://www.clinicaltrials.gov/ct2/show/NCT02713828; last access: 01.03.2017

[2] AstraZeneca. Assessment of the efficacy and safety of olaparib monotherapy versus physicians choice chemotherapy in the treatment of metastatic breast cancer patients with germline BRCA1/2 mutations. (OlympiAD). 2013. Online: https://www.clinicaltrials.gov/ct2/show/NCT02000622; last access: 01.03.2017

[3] AstraZeneca. Olaparib as adjuvant treatment in patients with germline BRCA mutated high risk HER2 negative primary breast cancer (OlympiA). 2017. Online: https://www.clinicaltrials.gov/ct2/show/NCT02032823; last access: 01.03.2017

[4] Althuis MD, Fergenbaum JH, Garcia-Closas M et al. Etiology of hormone receptor-defined breast cancer: a systematic review of the literature. Cancer Epidemiol Biomarkers Prev 2004; 13: 1558–1568

[5] Hess KR, Pusztai L, Buzdar AU et al. Estrogen receptors and distinct patterns of breast cancer relapse. Breast Cancer Res Treat 2003; 78: 105–118

[6] Yang XR, Chang-Claude J, Goode EL et al. Associations of breast cancer risk factors with tumor subtypes: a pooled analysis from the Breast Cancer Association Consortium studies. J Natl Cancer Inst 2011; 103: 250–263

[7] Rauh C, Gass P, Heusinger K et al. Association of molecular subtypes with breast cancer risk factors: a case-only analysis. Eur J Cancer Prev 2015; 24: 484–490

[8] Heusinger K, Jud SM, Haberle L et al. Association of mammographic density with hormone receptors in invasive breast cancers: results from a case-only study. Int J Cancer 2012; 131: 2643–2649

[9] Yaghjyan L, Colditz GA, Collins LC et al. Mammographic breast density and subsequent risk of breast cancer in postmenopausal women according to tumor characteristics. J Natl Cancer Inst 2011; 103: 1179–1189

[10] SP&A Application Laboratory Thermo Fisher Scientific Vantaa Finland. Rapid and non-invasive SNP determination of lactase persistence trait – application notes. 2014. Online: https://tools.thermofisher.com/content/sfs/brochures/SNP-Determination-of-Lactose-AppNote-EN.pdf; last access: 05.05.2017

[11] Couch FJ, Hart SN, Sharma P et al. Inherited mutations in 17 breast cancer susceptibility genes among a large triple-negative breast cancer cohort unselected for family history of breast cancer. J Clin Oncol 2015; 33: 304–311

[12] Fasching PA, Ekici AB, Wachter DL et al. Breast cancer risk – from genetics to molecular understanding of pathogenesis. Geburtsh Frauenheilk 2013; 73: 1228–1235

[13] Michailidou K, Hall P, Gonzalez-Neira A et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet 2013; 45: 353–361, 361e351–361e352

[14] Michailidou K, Beesley J, Lindstrom S et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat Genet 2015; 47: 373–380

[15] Purrington KS, Slager S, Eccles D et al. Genome-wide association study identifies 25 known breast cancer susceptibility loci as risk factors for triple-negative breast cancer. Carcinogenesis 2014; 35: 1012–1019

[16] Warren H, Dudbridge F, Fletcher O et al. 9q31.2-rs865686 as a susceptibility locus for estrogen receptor-positive breast cancer: evidence from the Breast Cancer Association Consortium. Cancer Epidemiol Biomarkers Prev 2012; 21: 1783–1791

[17] Stevens KN, Fredericksen Z, Vachon CM et al. 19p13.1 is a triple-negative-specific breast cancer susceptibility locus. Cancer Res 2012; 72: 1795–1803

[18] Stevens KN, Vachon CM, Lee AM et al. Common breast cancer susceptibility loci are associated with triple-negative breast cancer. Cancer Res 2011; 71: 6240–6249

[19] Garcia-Closas M, Couch FJ, Lindstrom S et al. Genome-wide association studies identify four ER negative-specific breast cancer risk loci. Nat Genet 2013; 45: 392–398

[20] Antoniou AC, Wang X, Fredericksen ZS et al. A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population. Nat Genet 2010; 42: 885–892

[21] Broeks A, Schmidt MK, Sherman ME et al. Low penetrance breast cancer susceptibility loci are associated with specific breast tumor subtypes: findings from the Breast Cancer Association Consortium. Hum Mol Genet 2011; 20: 3289–3303

[22] Heusinger K, Loehberg CR, Haeberle L et al. Mammographic density as a risk factor for breast cancer in a German case-control study. Eur J Cancer Prev 2011; 20: 1–8

[23] Rauh C, Hack CC, Haberle L et al. Percent mammographic density and dense area as risk factors for breast cancer. Geburtsh Frauenheilk 2012; 72: 727–733

[24] Haberle L, Wagner F, Fasching PA et al. Characterizing mammographic images by using generic texture features. Breast Cancer Res 2012; 14: R59

[25] French JD, Ghoussaini M, Edwards SL et al. Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers. Am J Hum Genet 2013; 92: 489–503

[26] Dunning AM, Healey CS, Baynes C et al. Association of ESR1 gene tagging SNPs with breast cancer risk. Hum Mol Genet 2009; 18: 1131–1139

[27] Schmidt MK, Hogervorst F, van Hien R et al. Age- and tumor subtype-specific breast cancer risk estimates for CHEK2*1100delC carriers. J Clin Oncol 2016; 34: 2750–2760

[28] Weischer M, Nordestgaard BG, Pharoah P et al. CHEK2*1100delC heterozygosity in women with breast cancer associated with early death, breast cancer-specific death, and increased risk of a second breast cancer. J Clin Oncol 2012; 30: 4308–4316

[29] Milne RL, Benitez J, Nevanlinna H et al. Risk of estrogen receptor-positive and -negative breast cancer and single-nucleotide polymorphism 2q35-rs13387042. J Natl Cancer Inst 2009; 101: 1012–1018

[30] Bojesen SE, Pooley KA, Johnatty SE et al. Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. Nat Genet 2013; 45: 371–384, 384e371–384e372

[31] Dunning AM, Michailidou K, Kuchenbaecker KB et al. Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170. Nat Genet 2016; 48: 374–386

[32] Ghoussaini M, Fletcher O, Michailidou K et al. Genome-wide association analysis identifies three new breast cancer susceptibility loci. Nat Genet 2012; 44: 312–318

[33] Haiman CA, Chen GK, Vachon CM et al. A common variant at the TERT-CLPTM1 L locus is associated with estrogen receptor-negative breast cancer. Nat Genet 2011; 43: 1210–1214

[34] Beckmann MW, Brucker C, Hanf V et al. Quality assured health care in certified breast centers and improvement of the prognosis of breast cancer patients. Onkologie 2011; 34: 362–367

[35] Easton DF, Pooley KA, Dunning AM et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 2007; 447: 1087–1093

[36] Azzato EM, Tyrer J, Fasching PA et al. Association between a germline OCA2 polymorphism at chromosome 15q13.1 and estrogen receptor-negative breast cancer survival. J Natl Cancer Inst 2010; 102: 650–662

[37] Fagerholm R, Hofstetter B, Tommiska J et al. NAD(P)H : quinone oxidoreductase 1 NQO1*2 genotype (P187S) is a strong prognostic and predictive factor in breast cancer. Nat Genet 2008; 40: 844–853

[38] Pirie A, Guo Q, Kraft P et al. Common germline polymorphisms associated with breast cancer-specific survival. Breast Cancer Res 2015; 17: 58

[39] Shu XO, Long J, Lu W et al. Novel genetic markers of breast cancer survival identified by a genome-wide association study. Cancer Res 2012; 72: 1182–1189

[40] Goldhirsch A, Wood WC, Gelber RD et al. Meeting highlights: updated international expert consensus on the primary therapy of early breast cancer. J Clin Oncol 2003; 21: 3357–3365

[41] Goldhirsch A, Glick JH, Gelber RD et al. Meeting highlights: international expert consensus on the primary therapy of early breast cancer 2005. Ann Oncol 2005; 16: 1569–1583

[42] Goldhirsch A, Wood WC, Gelber RD et al. Progress and promise: highlights of the international expert consensus on the primary therapy of early breast cancer 2007. Ann Oncol 2007; 18: 1133–1144

[43] Goldhirsch A, Ingle JN, Gelber RD et al. Thresholds for therapies: highlights of the St Gallen International Expert Consensus on the primary therapy of early breast cancer 2009. Ann Oncol 2009; 20: 1319–1329

[44] Sauter G, Lee J, Bartlett JM et al. Guidelines for human epidermal growth factor receptor 2 testing: biologic and methodologic considerations. J Clin Oncol 2009; 27: 1323–1333

[45] Salmen J, Neugebauer J, Fasching PA et al. Pooled analysis of the prognostic relevance of progesterone receptor status in five German cohort studies. Breast Cancer Res Treat 2014; 148: 143–151

[46] Harrell FE jr., Lee KL, Pollock BG. Regression models in clinical studies: determining relationships between predictors and response. J Natl Cancer Inst 1988; 80: 1198–1202

[47] Bovelstad HM, Nygard S, Storvold HL et al. Predicting survival from microarray data–a comparative study. Bioinformatics 2007; 23: 2080–2087

[48] Steyerberg EW, Borsboom GJ, van Houwelingen HC et al. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med 2004; 23: 2567–2586

[49] Sauerbrei W, Schumacher M. A bootstrap resampling procedure for model building: application to the Cox regression model. Stat Med 1992; 11: 2093–2109

[50] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 1996; 58: 267–288

[51] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001; 29: 1189–1232

[52] Bühlmann P, Hothorn T. Boosting algorithms: regularization, prediction and model fitting. Stat Sci 2007; 22: 477–505

[53] Boulesteix AL, Hothorn T. Testing the additional predictive value of high-dimensional molecular data. BMC Bioinformatics 2010; 11: 78

[54] Pencina MJ, D'Agostino RB sr., Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med 2011; 30: 11–21

[55] Wessels LF, Reinders MJ, Hart AA et al. A protocol for building and evaluating predictors of disease state based on microarray data. Bioinformatics 2005; 21: 3755–3762

[56] Häberle L, Fasching PA, Brehm B et al. Mammographic density is the main correlate of tumors detected on ultrasound but not on mammography. Int J Cancer 2016; 139: 1967–1974

[57] Kote-Jarai Z, Saunders EJ, Leongamornlert DA et al. Fine-mapping identifies multiple prostate cancer risk loci at 5p15, one of which associates with TERT expression. Hum Mol Genet 2013; 22: 2520–2528

[58] Meyer KB, O'Reilly M, Michailidou K et al. Fine-scale mapping of the FGFR2 breast cancer risk locus: putative functional variants differentially bind FOXA1 and E2F1. Am J Hum Genet 2013; 93: 1046–1060

[59] Garcia-Closas M, Hall P, Nevanlinna H et al. Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. PLoS Genet 2008; 4: e1000054

[60] Masson JY, Tarsounas MC, Stasiak AZ et al. Identification and purification of two distinct complexes containing the five RAD51 paralogs. Genes Dev 2001; 15: 3296–3307

[61] Thacker J. The RAD51 gene family, genetic instability and cancer. Cancer Lett 2005; 219: 125–135

[62] Orr N, Lemnrau A, Cooke R et al. Genome-wide association study identifies a common variant in RAD51B associated with male breast cancer risk. Nat Genet 2012; 44: 1182–1184

[63] Eeles RA, Olama AA, Benlloch S et al. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. Nat Genet 2013; 45: 385–391, 391e381–391e382

[64] Couch FJ, Wang X, McGuffog L et al. Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. PLoS Genet 2013; 9: e1003212

[65] Lee PS, Fang J, Jessop L et al. RAD51B activity and cell cycle regulation in response to DNA damage in breast cancer cell lines. Breast Cancer (Auckl) 2014; 8: 135–144

[66] Lambrechts D, Truong T, Justenhoven C et al. 11q13 is a susceptibility locus for hormone receptor positive breast cancer. Hum Mutat 2012; 33: 1123–1132

[67] Turnbull C, Ahmed S, Morrison J et al. Genome-wide association study identifies five new breast cancer susceptibility loci. Nat Genet 2010; 42: 504–507

[68] Finn RS, Aleshin A, Slamon DJ. Targeting the cyclin-dependent kinases (CDK) 4/6 in estrogen receptor-positive breast cancers. Breast Cancer Res 2016; 18: 17

[69] Ein-Dor L, Kela I, Getz G et al. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics 2005; 21: 171–178

[70] Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci U S A 2002; 99: 6562–6566

[71] Schild RL, Maringa M, Siemer J et al. Weight estimation by three-dimensional ultrasound imaging in the small fetus. Ultrasound Obstet Gynecol 2008; 32: 168–175

[72] Hepp T, Schmid M, Gefeller O et al. Approaches to regularized regression–a comparison between gradient boosting and the lasso. Methods Inf Med 2016; 55: 422–430

[73] Rauh C, Schuetz F, Rack B et al. Hormone therapy and its effect on the prognosis in breast cancer patients. Geburtsh Frauenheilk 2015; 75: 588–596

[74] Häberle L, Wagner F, Fasching PA et al. Characterizing mammographic images using generic texture features. Breast Cancer Res 2012; 14: R59

[75] Moons KG, Kengne AP, Woodward M et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart 2012; 98: 683–690

[76] Gerds TA, Cai T, Schumacher M. The performance of risk prediction models. Biom J 2008; 50: 457–479

[77] Wang TJ, Gona P, Larson MG et al. Multiple biomarkers for the prediction of first major cardiovascular events and death. N Engl J Med 2006; 355: 2631–2639

[78] Pencina MJ, D'Agostino RB sr., D'Agostino RB jr. et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008; 27: 157–172; discussion 207–212

[79] Thomas G, Jacobs KB, Kraft P et al. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). Nat Genet 2009; 41: 579–584

[80] Cox A, Dunning AM, Garcia-Closas M et al. A common coding variant in CASP8 is associated with breast cancer risk. Nat Genet 2007; 39: 352–358

[81] Stacey SN, Manolescu A, Sulem P et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 2007; 39: 865–869

[82] Couch FJ, Kuchenbaecker KB, Michailidou K et al. Identification of four novel susceptibility loci for oestrogen receptor negative breast cancer. Nat Commun 2016; 7: 11375

[83] Ahmed S, Thomas G, Ghoussaini M et al. Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. Nat Genet 2009; 41: 585–590

[84] Milne RL, Burwinkel B, Michailidou K et al. Common non-synonymous SNPs associated with breast cancer susceptibility: findings from the Breast Cancer Association Consortium. Hum Mol Genet 2014; 23: 6096–6111

[85] Stacey SN, Manolescu A, Sulem P et al. Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 2008; 40: 703–706

[86] Johnatty SE, Beesley J, Chen X et al. Evaluation of candidate stromal epithelial cross-talk genes identifies association between risk of serous ovarian cancer and TERT, a cancer susceptibility "hot-spot". PLoS Genet 2010; 6: e1001016

[87] Wang Y, McKay JD, Rafnar T et al. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. Nat Genet 2014; 46: 736–741

[88] Zheng W, Long J, Gao YT et al. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nat Genet 2009; 41: 324–328

[89] Fletcher O, Johnson N, Orr N et al. Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study. J Natl Cancer Inst 2011; 103: 425–435