

A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance

Catherine Meads · Ikhlmaq Ahmed ·
Richard D. Riley

Received: 30 September 2011 / Accepted: 1 October 2011 / Published online: 22 October 2011
© Springer Science+Business Media, LLC. 2011

Abstract A risk prediction model is a statistical tool for estimating the probability that a currently healthy individual with specific risk factors will develop a condition in the future such as breast cancer. Reliably accurate prediction models can inform future disease burdens, health policies and individual decisions. Breast cancer prediction models containing modifiable risk factors, such as alcohol consumption, BMI or weight, condom use, exogenous hormone use and physical activity, are of particular interest to women who might be considering how to reduce their risk of breast cancer and clinicians developing health policies to reduce population incidence rates. We performed a systematic review to identify and evaluate the performance of prediction models for breast cancer that contain modifiable factors. A protocol was developed and a sensitive search in databases including MEDLINE and EMBASE was conducted in June 2010. Extensive use was made of reference lists. Included were any articles proposing or validating a breast cancer prediction model in a general female population, with no language restrictions. Duplicate data extraction and quality assessment were conducted. Results were summarised qualitatively, and where possible meta-analysis of model performance statistics was undertaken. The systematic review found 17 breast cancer models, each containing a different but often overlapping

set of modifiable and other risk factors, combined with an estimated baseline risk that was also often different. Quality of reporting was generally poor, with characteristics of included participants and fitted model results often missing. Only four models received independent validation in external data, most notably the ‘Gail 2’ model with 12 validations. None of the models demonstrated consistently outstanding ability to accurately discriminate between those who did and those who did not develop breast cancer. For example, random-effects meta-analyses of the performance of the ‘Gail 2’ model showed the average *C* statistic was 0.63 (95% CI 0.59–0.67), and the expected/observed ratio of events varied considerably across studies (95% prediction interval for *E/O* ratio when the model was applied in practice was 0.75–1.19). There is a need for models with better predictive performance but, given the large amount of work already conducted, further improvement of existing models based on conventional risk factors is perhaps unlikely. Research to identify new risk factors with large additionally predictive ability is therefore needed, alongside clearer reporting and continual validation of new models as they develop.

Keywords Breast cancer · Systematic review · Prediction models

C. Meads (✉)
Centre for Primary Care and Public Health, Barts and The
London School of Medicine and Dentistry, Queen Mary
University of London, Yvonne Carter Building, 58 Turner St,
Whitechapel, London E1 2AB, UK
e-mail: c.meads@qmul.ac.uk

I. Ahmed · R. D. Riley
Unit of Public Health, Epidemiology and Biostatistics,
University of Birmingham, B15 2TT Birmingham, UK

Background

A risk prediction model is a statistical tool for estimating the probability that a currently healthy individual with specific risk factors (e.g. age, menopausal status) will develop a future condition, such as breast cancer, within a certain time period (such as within 5 years or lifetime). Risk models combine the baseline risk of developing the

condition with an individual's risk score, i.e. a score derived from their set of risk factor values multiplied by the 'beta' weights (e.g. log odds ratios) associated with these factors, as estimated from a statistical equation. The baseline risk of the condition represents the underlying population risk for patients whose risk factor values are all zero (or 'not present'), and this is usually estimated from a prospective population-based cohort study. The risk-score component shows how much the baseline risk is multiplied for increasing values of the risk factors, and may also be estimated using a cohort study or, for rare conditions, a case-control study. The two main statistical models used to identify important risk factors and to estimate their associated beta weights are logistic regression and Cox proportional hazards regression. The variables in the model can be any combination of environmental, behavioural, genetic or psychological attributes of the person. As well as estimating risk estimates for specific individuals, risk predictions model can also make a population-based estimate of risk by using average risk factor values from the population.

After a risk prediction model has been developed in a sample from a population, it then needs to be validated in further independent samples from the same population, and indeed within samples from different populations to ensure that it is reliable and generalisable. Frequently the predictive accuracy of the model is not as good in the validation sample as the original sample, and so adjustments are made leading to new or modified models being gradually developed over time. Even if the same risk factor variables are included, their beta weights may be changed which would then constitute a different model, as the risk score would then change. The performance of prediction models may also vary according to the population they are applied to, so that a model may have good accuracy in a high risk population and not in a low risk population and vice versa. There are a wide variety of ways to describe the performance of predictive tests. Steyerberg et al. [1] describe a number of these, and many of the statistical measures used are well-known from diagnostic test studies, such as sensitivity, specificity and the AUC (area under the receiver operating characteristic curve (ROC)) [2].

In practice, only some of these statistics are reported in modelling and validation articles for risk prediction models, and the two most common are the *E/O* statistic and the *C* statistic, interpreted as follows:

- The *E/O* statistic measures the *calibration* performance of the model. It compares expected (*E*) numbers to observed (*O*) numbers of events, so a well fitting model should have the number close to 1. A number lower underestimates the incidence of the condition whereas a number higher overestimates the incidence. Often

E/O statistics are presented for deciles of the population defined by predicted risk, to see whether *E/O* is close to 1 in all deciles or not.

- The *C* statistic measures the *discrimination* performance of the model. It gives the proportion of randomly chosen pairs from the sample (i.e. a person with the condition paired with one without it), where the person with the condition has a higher predicted risk than the one without. A *C* statistic of 0.5 is equivalent to no discrimination between people who develop the condition and those who do not, whereas 1.0 indicates perfect discrimination. The AUC and the *C* statistic are the same.

Such performance statistics can be used to compare competing risk prediction models. Ideally one or more primary studies would need to be done that used the same data to run all of the different models and discover which one was superior in terms of calibration and discrimination. Such studies are, however, rare and so systematic reviews of the literature to identify, extract and synthesise performance statistics for each model across multiple studies are important.

Risk models in breast cancer

A large number of risk prediction models have been developed that have looked at a variety of different risk factors for developing breast cancer. The most well-known is the Gail model but there are numerous others such as the Claus model, the Tyrer-Cuzick model and the Jonker model [3]. These models for the prediction of breast cancer incidence should not be confused with models that predict recurrence or mortality after the initial breast cancer incidence; these are known as prognostic models rather than risk prediction models [4]. There are numerous breast cancer prognostic models and a recent health technology assessment reviewed 61 prognostic models of single or multiple factors [5, 6].

Family history is an important risk factor for breast cancer and is usually included in these prediction models. Some models explore breast cancer incidence in mothers and sisters as risk factors, whilst others also look at family history of ovarian cancer. Where the models only looks at family history, the variables investigated are not modifiable, as opposed to other models which include variables such as alcohol consumption, BMI or weight, condom use, exogenous hormone use and physical activity which are all potentially modifiable. Many women are interested in whether they can reduce their risk of breast cancer so it would be useful to know which modifiable variables are included within existing prediction models and, if altered, the extent to which they might reduce their risk of breast

cancer. Clinicians and health policy makers may also aim to lower the population rate of breast cancer by public health interventions intended to reduce modifiable risk factors.

Scoping the literature prior to our review, we found two relevant systematic or semi-systematic reviews of incidence risk prediction models for breast cancer. One, published in 2007, searched Medline and PubMed only and did not assess breast cancer risk prediction models in depth [7]. A systematic review of breast cancer risk assessment models [3] published in 2009 listed seven models, of which four included family history variables only (BOADICEA, Claus, Claus2, Jonker) one was a modification of the Claus model (from a article by van Asperen from 2004) and the remaining two were the Gail and the Tyrer-Cuzick models, both of which do include modifiable risk factors. The systematic review did not synthesise model performance statistics across multiple studies, and did not fully cover the possible list of models available as several other models have been published [8, 9]. Therefore, there is no currently published systematic review assessment of the most accurate breast cancer incidence prediction model to use in the UK population or similar. In this article, we report a systematic review identifying articles developing or validating breast cancer incidence risk prediction models that contain at least one modifiable risk factor. The aim is to qualitatively summarise the models and the risk factors they contain, and to collate and meta-analyse model performance statistics across studies, to allow the performance of each model to be compared.

Methods

A protocol was developed and a scoping search undertaken in November 2009. For the full systematic review the following databases were searched from inception to June 2010: Cochrane library (CDSR, CENTRAL, HTA, DARE, NHSEED), MEDLINE, EMBASE, CAB Abstracts and PsychINFO. The following search terms were used: breast cancer and prediction or risk model as index terms and text words as this was found to maximise the number of relevant citations. A sensitive search strategy was used as there is no specific MESH term and database indexing of prediction model studies is relatively poor. Extensive use was made of reference lists in the systematic reviews and primary studies to find further studies for inclusion.

Inclusion criteria were any studies developing and/or validating a breast cancer risk prediction model for the general female population using multiple variables, at least one of which was a modifiable risk factor. There was no restriction on study type; for example, studies may use data from cohort studies, prospectively or retrospectively

recruited, where some healthy women at the start went on to develop breast cancer, or they may use data from case control studies where cases had breast cancer and controls did not, with risk factors ascertained from existing records. There were no language restrictions. Excluded were breast cancer in men, women who already had breast cancer or benign breast pathology when recruited and studies in high risk groups of women such as with specific genetic mutations or who have close family relatives with breast cancer. Models investigating single risk factor results such as mammography, assessing genetic risk factors only including carrier status of mutations, assessing invasive techniques such as biopsies only, assessing family history of breast or ovarian cancer only and predicting genetic mutations rather than cancer were excluded. Screening and early detection studies were excluded, as were models published more than 25 years previously (i.e. before 1985).

All identified citations (titles ± abstracts) were screened by one reviewer for inclusion and 10% were checked by a second reviewer with no discrepancies. Full articles were ordered for all included or possibly included citations and screened for inclusion. All data extraction was conducted by two reviewers independently and disagreements were resolved by discussion with a senior statistician. There are no specific quality assessment checklists for prediction modelling studies but there is a list of criteria recently published by Altman [6]. This was for prognostic rather than predictive studies but the issues are similar. The quality factors he discusses that are relevant to prediction models are listed below.

- Study design—cohort usually of higher quality due to risk factors being recorded at baseline and time of event used.
- Patient sample—prospective data collection with characteristics of patients recorded.
- Sample size—the power of the study depends on the number of events so if a general population sample is used it will need large numbers or long follow up or both. It was suggested that there needs to be ten times the number of events to the number of variables studied.
- Incomplete data, missing or losses to follow up—these are difficult problems that reduce power and probably result in bias. Completeness of data should be reported.
- Variables—use of continuous variables on their original scale is better than categorisation of a continuous variable into, for example, low and high risk groups, as categorisation of a continuous variable reduces power.
- Presentation—whether full presentation of the model is given including all of the variables, their beta weights, and the baseline risk estimate; this is important to enable the model to be applied in practice.

- Validation—ideally the models need to have validation from an independent sample rather than looking at model performance in the same sample used to develop the model, or even when splitting the sample in order to use half to develop the model and the other half to validate it.

For each study identified by our review, we extracted information relating to each of these criteria and noted when pertinent information (e.g. missing data, model beta estimates provided, etc.) was not reported. Study characteristics and results were extracted, tabulated and assessed qualitatively and graphically where appropriate. Results were interpreted in light of methodological strengths and weaknesses identified in quality assessment. For validation studies, any metric of model performance (such as the *E/O* statistic and the *C* statistic) were extracted alongside their uncertainty (e.g. their confidence interval or standard error). Where observed over expected (*O/E*) rates were given in articles, these have been reported in the summary tables but converted to *E/O* to give consistency in the meta-analysis. Where multiple studies reported the same performance statistic for a model, a random-effects meta-analysis was performed, using the DerSimonian and Laird method via STATA software version 11 [10, 11], to summarise model performance. This approach estimates the average performance of the model, the between-study heterogeneity in model performance, and a 95% prediction interval for the model performance when it is applied in a single population setting [12].

Results

From the database searches 7,317 references were found of which 1,265 were duplicates. The flow of articles is shown in Fig. 1. The following studies were found:

- Six studies describing the development of a new prediction model (Arne 2009, Colditz 2000, Cook 2009, Gail 1989, Tyrer 2004 and Wacholder 2010).
- Nine studies validating one or more prediction models in a new sample from a potentially different population ((Gail 1 or 2 model validation: Amir 2003 (also validated Tyrer 2004), Bondy 1994, Costantino 1999, Rockhill 2001, Speigelman 1994, Schonfeld 2010, Ulusoy 2010) (Rosner 1994 model validation: Rockhill 2003 [13] (also validated Colditz 2000), Viallon 2009 [14])).
- Eleven studies both describing and validating one or more models (Barlow 2006 (validated Gail 2), Boyle 2004 (validated Gail 2), Chen 2006 (validated Gail 2), Decarli 2006 (validated Gail 2), Gail 2007 (validated Gail 2), Novotny 2006 (validated Gail 2), Rosner 1994 (validated Pike—but this model excluded as published

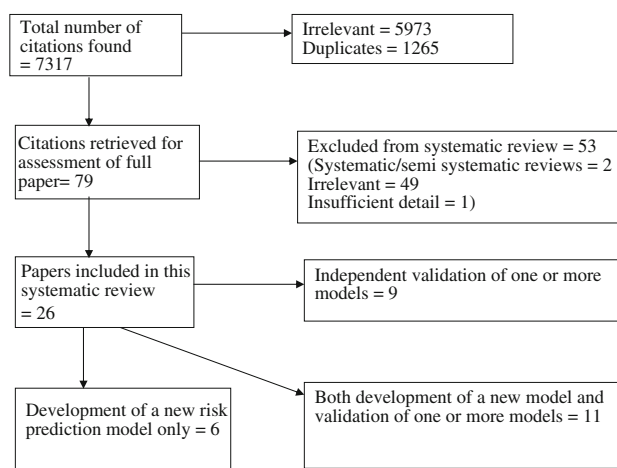


Fig. 1 PRISMA diagram for systematic review

in 1983), Rosner 1996 (validated Pike—ditto), Rosner 2008 (validated Colditz), Tice 2005 (validated Gail 2), Tice 2008 (validated Gail 2)).

Studies describing a new risk prediction model

In total, 17 studies developing a new risk prediction model were found and several of these described more than one version, and new models were often a modification of previously proposed models (e.g. with the addition of one or more risk factors). The models and their reporting are described in Table 1. None of the 17 prediction model studies gave a justification for the sample size used, though most included large numbers of participants. Fifteen of the 17 articles stated the number of eligible patients for model derivation; seven gave the number of events per variable; eight summarised the sample characteristics (e.g. mean age, proportion post-menopausal) in a table; nine stated whether there was missing risk factor data for some participants; and fifteen stated how they handled continuous variables (i.e. whether kept continuous or categorised). Only six reported the full specification of the final developed model(s), including parameter values (beta weights) and their standard errors or 95% confidence intervals for all included variables. Of those 11 articles that did not report the full model, eight alternatively gave beta weights for a partial set of variables and/or gave transformed beta weights (e.g. odds ratios or risk ratios) for some or all variables. The variables used in each of the prediction models are shown in Table 2. The modifiable risk factors that were included in one or more models were alcohol consumption, breast biopsy number, BMI or weight, condom use, exogenous hormone use (HRT, contraceptive pill), and physical activity. The most commonly included risk factors in the models were age, Age at first live birth

Table 1 Risk prediction model reporting and analysis characteristics

Article	Description of key aspects of study design		Sample characteristics	Data quality	Handling of continuous variables	Presentation of model	
	Number of eligible patients given	Number of events per variable				Were they summarised in a table (e.g. mean age, proportion males, ...)	Missing data for each variable mentioned/stated in the article?
Arne 2009	No	Not stated	No	No	Not given	No	No
Barlow 2006	Yes, 1,007,600	Stated	Yes	Yes	Categorised (test for trend across categories of a continuous factor performed)	No	Yes; ORs and CIs for each category of each variable relative to the reference category
Boyle 2004	Yes, 5,157	Not stated	Yes	No	Categorised	No (only some of the final model variables are given)	Yes; ORs and CIs given for a partial set of the included variables; score chart provided
Chen 2006	Yes, 284,780	Not stated	No	Yes	Categorised	No	Yes (variable names with coefficients, but no CIs)
Colditz 2000	Yes, 58,520	Not stated	No	No	Kept continuous	Yes	
Cook 2009	Yes, 45,281	Not stated	Yes	Yes, as complete data available	Kept continuous	Yes	
Decarli 2006	Yes, 5,157	Stated	Yes	Yes; patients with missing data excluded	Categorised	No	Yes; ORs and CI given for variables
Gail 1989	Yes, 5,998	Stated	No	No	Continuous but age categorised	YES (estimates and their standard errors given)	
Gail 2007	Yes, 3,254	Stated	Yes	Yes (NB they had complete data)	Categorised	YES (estimates and their standard errors given)	No
Novotny 2006	Yes, 4,598	Not stated	No	No	Categorised	No	Yes; ORs and parameter estimates given, but no standard errors or CIs
Rosner 1994	Yes, 91,523	Not stated	No	Yes: patients with missing data excluded	Continuous	No	Yes; some parameter estimates and some CIs given
Rosner 1996	Yes, 89,132	Not stated	No	No	Continuous	Yes	
Rosner 2008	Yes, 59,812	Not stated	No	No	Continuous	Yes	
Tice 2005	Yes, 81,777	Stated	Yes	Yes	Categorised, as in the Gail model	No	Yes; some parameter estimates and CIs given
Tice 2008	Yes, 1,095,484	Stated	Yes	Yes	A mixture of continuous and categorisation used	No	No
Tyrer 2004	No	Not stated	No	No	Not given	No	No
Wacholder 2010	Yes, 11,588	Stated	Yes	Yes	Categorised	No	Yes; ORs and CIs for the variables

Table 2 Breast cancer risk models comparison of factors

Risk factors (modifiable factors in bold)	Arne 2009	Barlow 2006 pre-menopausal	Barlow 2006 post-menopausal	Boyle 2004	Chen 2006	Colditz 2000	Cook 2009	Decarli 2006	Gail 1989
Age	Y	Y	Y	Y		Y	Y	Y	Y
Age at menarche				Y		Y	Y	Y	Y
Age at first live birth and/or age at subsequent births			Y	Y	Y	Y	Y		Y
Age at menopause						Y	Y		
Alcohol consumption				Y		Y	Y		
Atypical hyperplasia/benign breast disease			Y			Y	y		
Breast density		Y	Y		Y				
Birth history/parity	Y					Y			
Birth index						Y	Y		
Breast biopsy number		Y	Y		Y			Y	Y
BMI or weight			Y	Y	Y	Y	Y		
Condom use	Y								
Ethnicity			Y						
Exogenous hormone use (pill, HRT)			Y	Y		Y	Y		
Family history of breast cancer	Y	Y	Y	Y	Y	Y	Y	Y	Y
Family history of any cancer									
Height						Y	Y		
Physical activity				Y					
Reproductive age period	Y								
Surgical menopause			Y			Y	Y		
Risk factors (modifiable factors in bold)	Gail 2007	Novotny 2006	Rosner 1994	Rosner & Colditz 1996	Rosner 2008	Tice 2005	Tice 2008	Tyrer-Cuzick 2004	Wacholder 2010 models ^a
Age	Y	Y	Y	Y	Y	Y	Y	Y	1, 2, 3
Age at menarche	Y	Y	Y	Y	Y	Y		Y	1, 2, 3
Age at first live birth and/or age at subsequent births	Y	Y	Y	Y	Y	Y		Y	1, 2, 3
Age at menopause			Y	Y	Y			Y	
Alcohol consumption					Y				
Atypical hyperplasia/benign breast disease		Y			Y			Y	
Breast density						Y	Y		
Birth history/parity					Y			Y	
Birth index									
Breast biopsy number	Y	Y		Y		Y	Y		1, 2, 3
BMI or weight		Y			Y			Y	
Condom use									
Ethnicity							Y		
Exogenous hormone use (pill, HRT)					Y				
Family history of breast cancer	Y	Y			Y	Y	Y		1, 2, 3
Family history of any cancer		Y							
Height					Y			Y	
Number of contraceptives		Y							
Physical activity									
Reproductive age period									
Surgical menopause					Y				

^a Five models were presented in Wacholder 2010 but only three had modifiable risk factors and have been presented here

and/or age at subsequent births and family history of breast cancer. The risk factors only included in one model were condom use, family history of any cancer, physical activity and reproductive age period.

The only models with independent validations were Colditz 2000, Gail 1 (1989) and 2 (1999), Rosner and Colditz 1994 and Tyrer and Cusick 2004. Each of the 17 studies developing a new risk prediction models are now briefly summarised.

The study by Arne (2009) [15] was conducted in Macedonia but seems to have used an incidence database from the USA. No details are given of the database itself but a book reference is provided. The article is orientated around establishing use of condoms as a risk factor for breast cancer. One hypothetical example of a 54-year-old woman gives an estimated breast cancer risk of 62.7%. No independent validation of the model is available.

The article by Barlow et al. [16] developed two breast cancer models, one for pre-menopausal women and one for post-menopausal women. The database used was developed from seven US registries of women aged 35–84 who were undergoing very regular mammography (yearly or 18-monthly). Given that mammography is a known risk factor for breast cancer [17], the incidence in these cohorts is likely to be higher than in the UK where mammography is conducted routinely at 3-yearly intervals in women aged over 50.

Boyle and colleagues [18] developed a model based on an Italian case control study with the breast cancer cases aged between 23 and 74 years (median 55) and the controls aged 20–74 admitted to hospitals in the same catchment area with acute conditions. The model was a modification of the Gail 2 model (see below). The validation population was taken from an Italian RCT for adjuvant breast cancer treatment with tamoxifen but the women had all had hysterectomies. It was unclear why they had all had hysterectomies and it was also unclear as to why this group were chosen to validate the model because, as they remark in the article, the risk of this cohort developing breast cancer is around 20% lower than the general population.

The study by Chen et al. [19] is also a modification of the Gail 2 model and compares the new model to the Gail 2 model in a US population. The article refers to three unpublished articles by Chen and colleagues and as these articles are unavailable, it makes assessment of the presented model problematical. The way the article is written, with large numbers of abbreviations, equations and assumptions that the reader will understand the mathematical notation, makes the article very difficult to interpret for practitioners wishing to implement prediction models for breast cancer.

The study by Colditz and Rosner [20] was a continuation of previous modelling work done by Rosner et al.

(see below) [21], but they present a different model. The population was a cohort of 121,700 nurses from the USA but only 58,520 were used, the remainder being excluded for a variety of reasons including data inconsistencies around births and parity, incomplete data on height, weight and birth control pill use, and having had a surgical menopause. There is a further article by the same authors (Rosner et al. [9]) describing further modifications of the model with reference to predicting oestrogen-positive breast cancer, but using the same population as the study by Colditz and Rosner [20].

The article by Cook et al. [17] had Rosner and Colditz as co-workers and also used the same population as the Colditz and Rosner (2000) model [20], but the Cook et al. (2009) model contained a different set of risk factors to the Colditz and Rosner (2000) model and the modifications by Rosner et al. [9]. Some of the article looks at predictors for mammography but this is not relevant here. The article presented beta-weight results by individual risk factors from the total model which makes it easier to see the relative importance of the different factors.

The article by Decarli et al. [22] developed two models, one using cancer data from a case control study and the other from a cancer registry, both models being based on a European cohort study—the Florence EPIC Cohort study and using variables from the Gail 2 model for invasive breast cancer (see below). Statistical methods and equations are presented in the text with the assumption that the reader will understand the mathematical notation.

The Gail model [23] was one of the earliest models and has undergone considerable development since the original article from 1989. The original model (Gail 1) used data from a case control study to estimate the beta weights (i.e. log odds ratios) for the risk factor variables. These were then combined with the baseline risk estimated using data from the BCDDP Cohort (a population of white women from USA), to allow individualised probabilities to be calculated.

A development of the Gail model, described here as the Gail 2 model, was published by Anderson and colleagues in 1992 as a technical report so is not universally available but is referred to in a article by Costantino et al. [24]. This model is for predicting invasive breast cancer only, and the baseline risk is estimated using a population of mixed ethnicity (The SEER database) available from the placebo arm of a RCT on adjuvant tamoxifen treatment. In this population the original Gail model (model 1) underestimated the risk of breast cancer ($E/O = 0.84$ (0.73–0.97)), particularly in the older women aged over 60 ($E/O = 0.66$ (0.52–0.86)). The Gail 2 model was used to establish eligibility for tamoxifen treatment by estimating baseline risk of breast cancer and is available on the internet (see <http://www.cancer.gov/bcrisktool/Default.aspx>). There

were several subsequent validations found for the Gail 1 and Gail 2 models (see below).

The article by Gail and colleagues 2007 [25] is a further development of the modelling initiative where another model was developed for invasive breast cancer risk in Black women using a case control study for relative and attributable risks, and the baseline risks estimated from Black women's data in a population cohort study (NCI SEER) and from national mortality data.

Novotny et al. [26] developed two Czech models using the same statistical approach as the Gail 2 model, including the same risk factor variables and four additional ones (number of relatives with any cancer, breast inflammation, BMI and number of conceptions) for the latter. The beta weights were estimated from a case control study in the Czech population.

The study by Rosner et al. [21] developed a model known as the Rosner and Colditz model which was a development of an early model by Pike and colleagues from 1983 but with an additional factor allowing for the influence of more than one birth. They also used a US Nurses cohort study for estimating their model parameters, whereas the Pike model used a US population cancer survey and case control studies. One modification study was found for the Rosner and Colditz model by Rosner and Colditz [27] which introduced a term to allow for a transient increase in risk with first pregnancy to increase with age at first pregnancy. This was because they noticed that the Rosner and Colditz model tended to overestimate incidence of breast cancer in younger women who had their babies young and underestimate incidence in older women who had their babies later.

Rosner et al. [9] developed another model, using their model from 1994, which evaluated whether oestradiol levels in post-menopausal women was a sufficiently important variable to merit inclusion into the model. Only oestrogen-positive breast cancers were considered. They concluded that it gave predictive value in addition to lifestyle factors.

Tice et al. [28] developed a model using a prospective cohort study from USA and included the variable of mammographic breast density with other variables from the Gail 2 model, with parameters estimated using Cox proportional hazards model. A new model was developed by Tice and colleagues (2008) [8] which incorporated mammographic breast density into a proportional hazards model. They used a population of US women from seven mammography registries (the same as those used to develop the Barlow models (see above)). They concluded that “the model has only modest ability to discriminate between women who will develop breast cancer and those who will not”.

The Tyrer and Cuzick model was developed in the UK [29] and is based on a Bayesian statistical analysis. The

population used was from the International breast intervention study (IBIS) and UK national statistics (Cancer Registry statistics from 1994). It combines extensive family history, atypical hyperplasia (benign breast disease) and endogenous oestrogen exposure.

The study by Wacholder et al. [30] was mainly orientated around evaluating the predictive power of genetic variants on the incidence of breast cancer. Five models were described, but only the three which incorporated modifiable risk factors are included here, and these were modifications of the Gail 2 model. Interestingly, the addition of the genetic components to the models only added a small increase in predictive power of between 2 and 4%.

Validations of prediction models

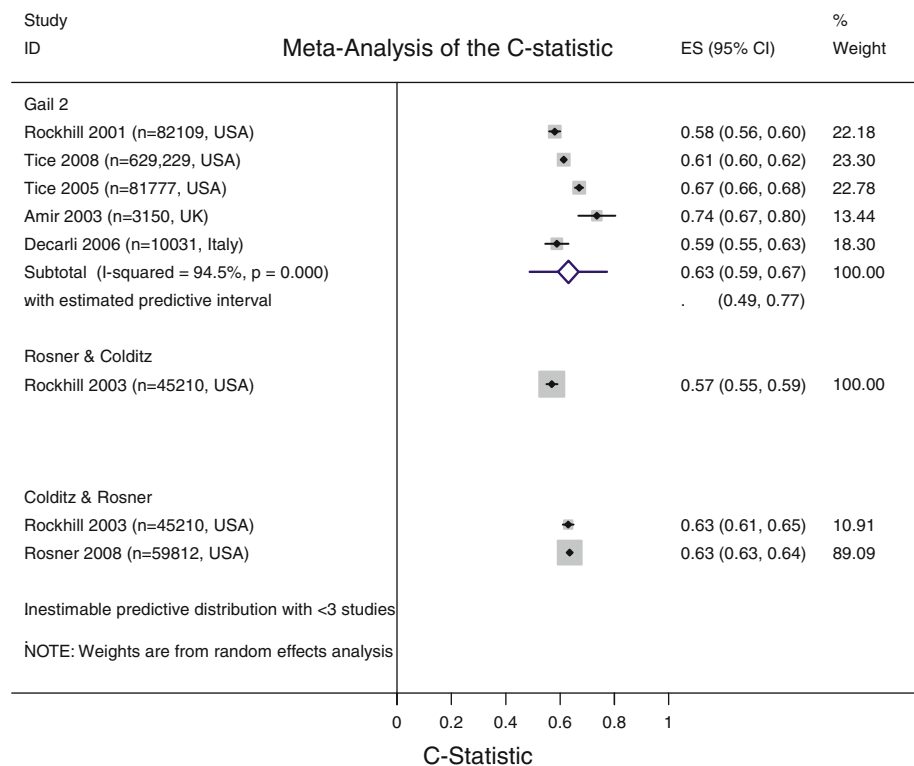
The only models listed above with independent validations in subsequent articles were Colditz 2000, Gail 1 and 2, Rosner and Colditz 1994 and Tyrer and Cuzick 2004. Validation study characteristics and results are shown in Table 3. There were four articles with independent validations of the Gail 1 model (Bondy 1994 [31], Costantino 1999 [32], Novotny 2006 [26], Spiegelman 2004 [33]) and 12 articles with independent validations of the Gail 2 model (Amir 2003 [34], Barlow 2006 [16], Boyle 2004 [18], Chen 2006 [19], Costantino 1999 [32], Decarli 2006 [22], Gail 2007 [25], Rockhill 2001 [35], Schonfeld 2010 [36], Tice 2005 [28], Tice 2008 [8], Ulusoy 2010 [37]) [37]. Note that Costantino 1999 [32] evaluated both Gail models, and several of the articles presented more than one validation using different cohorts. The validation populations were mostly from USA, but there was also one study from each of Czech Republic, France, Great Britain, Italy and Turkey. The standard of presentation was variable and some were very hard to interpret because of assumptions that the reader will understand mathematical notation, for example.

We extracted performance statistics from each of these articles where available. Meta-analyses of the *C* statistic (Fig. 2) and the *E/O* ratio (Fig. 3) were conducted for each model that had the validation statistic extracted from two or more articles. As validation statistics were often not reported, only a few meta-analyses were possible and often they contained a small number of studies. In most analyses there was considerable between-study heterogeneity in the validation statistic; for example, for Gail 2 model the proportion of total variability that is due to between-study heterogeneity (I^2) was 94.5 and 92.5% in the meta-analyses of the *C* statistic and *E/O* ratio, respectively. Such heterogeneity is perhaps unsurprising given the variations in populations used for the validations.

In the two meta-analyses of the *C* statistic (Fig. 2), the average *C* statistic was 0.63 (95% CI: 0.59–0.67, based on

Table 3 Breast cancer incidence model validation statistics

	Model validated	<i>C</i> statistic (95% CI)	Other validation stats	Validation population	Comments
Amir 2003	Gail 2	0.735 (0.666–0.803)	<i>E/O</i> = 0.69 (0.54–0.90)	High risk hospital cases and controls from UK	Also evaluated three genetic models
	Tyrer-Cuzick	0.762 (0.700–0.824)	<i>E/O</i> = 1.09 (0.85–1.41)		
Barlow 2006	Gail 2	0.598	Not given	Mammogram registry in USA population	Differences in time intervals for cancer ascertainment meant validation unreliable
Bondy 1994	Gail 1		<i>O/E</i> = 0.76	High risk white women in Texas USA	Subgroup analysis based on American Cancer Society mammogram screening guidelines given
Boyle 2004	Gail 2	Not given	<i>O/E</i> = 0.89 (0.70–1.09)	RCT of adjuvant tamoxifen in USA, all had hysterectomies	2 validations, above with original USA data set, below with Italian registry dataset. Unclear if for pre- or post-menopausal women or both.
		0.582	<i>O/E</i> = 0.96 (0.75–1.16)		
Chen 2006	Gail 2	0.602	Not given	Unclear	
Costantino 1999	Gail 1	Not given	<i>E/O</i> = 0.84 (0.73–0.97)	Women at increased risk of breast cancer in USA RCT of adjuvant tamoxifen	Distinguished clearly between total breast cancer (Gail 1) and invasive breast cancer (Gail 2)
	Gail 2		<i>E/O</i> = 1.03 (0.88–1.21)		
Decarli 2006	Gail 2	0.588 (0.546–0.631)	<i>E/O</i> = 0.93 (0.81–1.08)	Italian case control and registry studies	–
Gail 2007	Gail 2	0.636 (0.617–0.655)	<i>O/E</i> = 1.08 (0.97–1.20)	Black women from USA	Recalculated <i>C</i> statistic and <i>O/E</i> from data in article
Novotny 2006	Gail 1	Not given	Not given	Mammogram registry in Czech population	Only parameter estimates (ORs) with no standard errors of CIs given
Rockhill 2001	Gail 2		<i>E/O</i> = 0.94 (0.89–0.99)	White nurses in USA	Subgroup analyses for high risk and mammogram in past year also given
Rockhill 2003	Rosner & Colditz Colditz & Rosner	Not given	<i>E/O</i> = 1.00 (0.93–1.07)	Nurses from USA	Validation on same cohort as original model but using different time ranges
			<i>E/O</i> = 1.01 (0.94–1.09)		
Rosner 1994	Pike 1983	Not given	Not given	Nurses from USA	Pike model excluded as before 1985
Rosner 1996	Pike 1983	Not given	Not given	Nurses from USA	Pike model excluded as before 1985
Schonfeld 2010	Gail 2	Not given	Early SEER <i>E/O</i> = 0.87 (0.85–0.89)	White postmenopausal women from USA (NIH-AARP study)	Split SEER cohort by date from 1983–1987 and 1995–2003 and validated using two different populations
			Late SEER <i>E/O</i> = 1.03 (1.00–1.05)		
			Early SEER <i>E/O</i> = 0.86 (0.82–0.90)	White postmenopausal women from USA (PCLO trial)	
			Late SEER <i>E/O</i> = 1.01 (0.97–1.06)		
Rosner 2008	Colditz & Rosner	0.635 (0.628–0.642)	Not given	Nurses from USA	Focus of article on oestrogen receptor-positive breast cancer
Speigelman 1994	Gail 1	Not given	<i>E/O</i> = 1.33 (1.28–1.39)	Nurses from USA	Overprediction attributed to higher baseline incidence rates of breast cancer
Tice 2005	Gail 2	0.67 (0.65–0.68)	Not given	Mammography register in USA	ROC curve symmetrical
Tice 2008	Gail 2	0.613 (0.604–0.622)	Not given	7 mammogram registries from USA	Some data missing so authors recommend interpretation with caution
Ulusoy 2010	Gail 2	Not given	Using cut off risk ≥ 1.67 sensitivity = 13.3%, specificity = 92%, PPV = 63%, NPV = 51.9%	Turkish cases and controls from one hospital	Small validation sample
Viallon 2009	Rosner and Colditz	Not given	<i>E/O</i> = 0.947 (0.912–0.982)	French teacher, spouses and employees	Most of article describes mathematical simulations

Fig. 2 Meta-analyses of the C statistics

five articles) for Gail 2 model and also 0.63 (95% CI: 0.63–0.64, based on two articles) for the Colditz and Rosner model, which indicates only moderate performance. For the meta-analyses of the *E/O* ratios (Fig. 3), the average *E/O* ratio was 0.95 (0.88–1.01) for the Gail 2 model. This suggests that, on average (across all populations), it appears to have good calibration (as the estimate is close to 1 and the 95% CI contains 1), however, there is considerable heterogeneity in calibration across studies. This can be seen by the wide 95% prediction interval of 0.75–1.19, which gives the range of possible *E/O* ratios when the model is applied in any given study population. Gail model 1 had similar heterogeneity (Fig. 3). The Rosner and Colditz model showed less heterogeneity, but this may be due to it only being validated in two articles.

One study (Amir 2003) [38] was found which assessed the performance of several models on the same data set from South Manchester, UK. The population of 4,536 women had been assessed in a hospital clinic for breast and other cancer risks, where they completed a comprehensive risk assessment questionnaire, so were a high risk sample. Of the full sample, 1,933 women were followed up by screening every 12–18 months (screened population), 1,217 were discharged to routine follow up, 20 had breast cancer at the start and 1,366 women were lost to follow up during the study so this was a relatively small sample (3,150 total and 1,933 screened). Cancer incidence was found from the cancer registry but it was unclear how

successful the linkage was. It is unclear how the models assessed were chosen but they included well-known breast cancer models (Gail, Claus, Ford, Tyrer-Cuzick models and Claus tables). Other models available at the time but not assessed in this study included Rosner and Colditz 1994 [9, 39]. Amir and colleagues investigated the total population and the screened population and concluded that the Tyrer-Cuzick model was the most accurate for this high risk sample (see Table 4). It is unclear whether the Tyrer-Cuzick model would also be the best predictive model for a general population sample.

Discussion

The systematic review found a total of 17 risk prediction models, containing at least one modifiable risk factor, that aimed to predict breast cancer in populations and in individual women. Some have been independently validated, most notably Gail 2 with 12 validations, but many have no validations at all. However, none of the models that have been validated in more than one dataset demonstrated consistently outstanding performance, in terms of calibration or discrimination.

The standard of reporting of the articles could also be improved, and it was difficult at times to find basic information such as the number of events, amount of missing data, and even the fitted model results; similar problems

Fig. 3 Meta-analyses of the *E/O* ratios

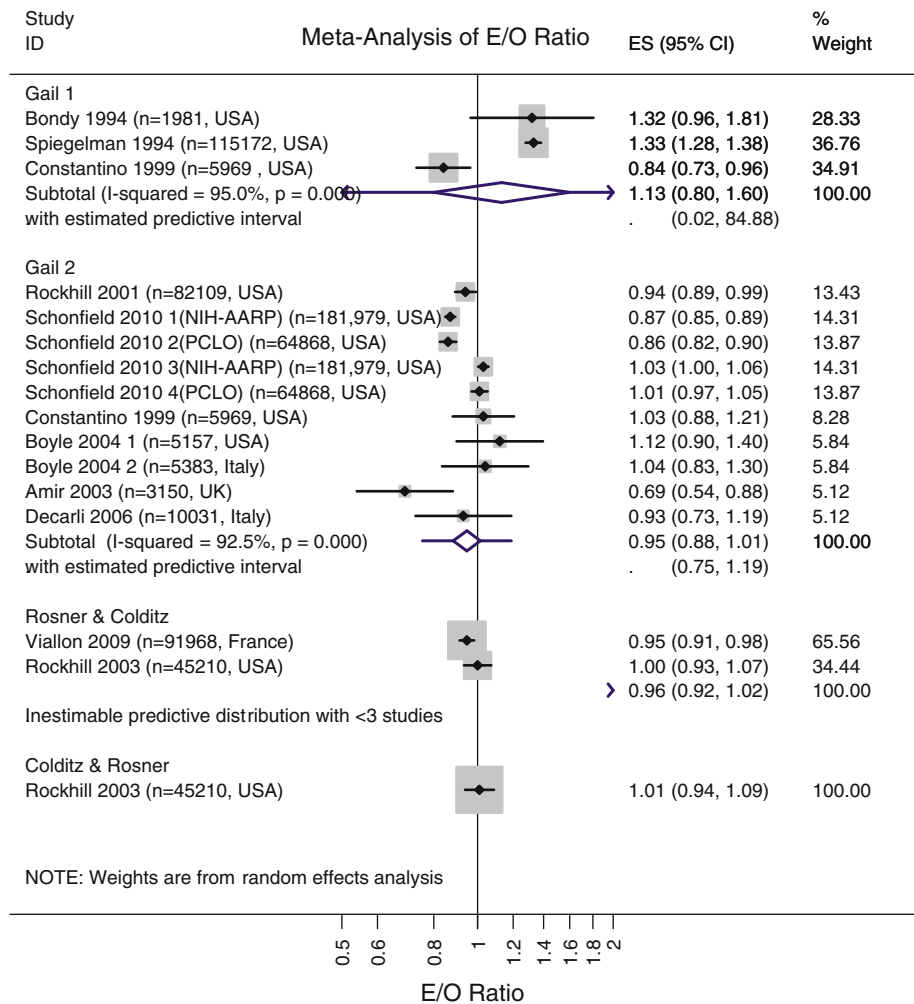


Table 4 Validations results from Amir 2003

	Gail 1	Gail 2	Tyrer-Cuzick
<i>E/O</i> (95% CI) (Screened population)	0.48 (0.37–0.64)	0.89 (0.68–1.20)	0.81 (0.62–1.08)
<i>C</i> statistic (ROC AUC, 95% CI)	0.735 (0.666–0.803)	0.727 (0.656–0.798)	0.762 (0.700–0.824)

have been found by prediction model articles in other fields [40, 41]. There are currently no accepted standards of reporting for risk prediction studies, unlike for RCTs (CONSORT guidelines) and for systematic reviews (PRISMA guidelines), though there is increasing interest in this area [42, 43]. Therefore, at the moment different authors are reporting different aspects of prediction models, and in varying ways, and single studies aiming to compare the performance of multiple models simultaneously are rare. This research focus generally seems to be on developing another new model, often by modifying an existing model, rather than seeking to validate existing models. There needs to be greater research effort into independent validation of promising models [44].

In this article, we have identified and attempted to compare existing models using a systematic review and meta-analyses. Though a few meta-analyses were possible, because of variable reporting and few validation studies there is insufficient information to distinguish the most accurate model, or indeed if any models consistently accurate enough for clinical practice. Further studies are needed; in particular primary studies that conduct a comparative valuation of all of the models on the same dataset(s), which has proved successful in other fields [45]. Only one study of this type was found (Amir) [34] but this evaluated a limited number of models.

Systematic reviews of prediction studies are rare [46] and there is no commonly accepted template. One strength

of this project is the comprehensiveness of the systematic review which included a sensitive search and extensive use of reference lists, so it is unlikely that any relevant studies will have been missed. Databases such as Medline and Embase do not have Medical Subject Heading (MESH) terms for prediction models so retrieval is not straightforward. The sensitive search used the term ‘prediction’ but in the published articles in the databases this term referred to prediction of breast cancer (incidence), prediction of breast cancer response to treatment and prediction of breast cancer recurrence. Prognostic and predictive terms were used synonymously in some articles and the predictive factors term was also used to describe the responsiveness of a tumour to specific treatments. It is acknowledged that screening for inclusions to systematic reviews by one reviewer only is likely to introduce biases, however, the second researcher screened 10% of the sample and found no discrepancies. On the other hand, double data extraction is a major strength of this review, as is the meta-analysis of results.

From our review it is difficult to recommend one model over another, and unfortunately no single model appears to perform consistently well. Also, there was no indication that any of the modifiable risk factors included in the models were sufficiently predictive of breast cancer to merit further investigation. From this it can be assumed that, at present, we have insufficient evidence to recommend that women attempt to modify any of these risk factors in order to reduce their individual risk of breast cancer. The main limitation of the models assessed is their relatively weak ability to predict risk of breast cancer, as found when their performance was evaluated independently in external data. Unfortunately, many promising models did not perform as well when considered in new data. This means that for individual women, many who have been identified as ‘low risk’ actually go on to develop breast cancer whereas many identified as ‘high risk’ do not [47]. It is possible that the weak predictive ability of the models stems from the fact that many of the included risk factors are common in society in women who never go on to develop breast cancer and are associated with relative risk of less than 10 whereas a good predictive factor would need a much higher relative risk [47]. Also, some of the more recent risk factor models were developed in order to determine eligibility for breast cancer adjuvant trials. Drug companies would have a vested interest in more women being eligible which may have biased the studies.

Conclusion

There is a need for further validation studies of existing risk prediction models for breast cancer, and for developing

improved breast cancer models with better predictive ability. None of the combinations of modifiable risk factors within the models was sufficiently powerful to predict breast cancer. Given the large amount of work already conducted on existing breast cancer models, it is uncertain whether it is possible to improve the currently available ones. Research is therefore also needed to identify new risk factors with high associated relative risk that add large predictive ability over and above currently used factors [48, 49].

Acknowledgments We would like to thank Dr Gill Lawrence for assistance and supervision at the beginning of this project.

Conflicts of interest None.

References

1. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N et al (2010) Assessing the performance of predictive models. *Epidemiology* 21(1):128–138
2. Meads CA, Cnossen JS, Meher S, Juarez-Garcia A, ter Riet G, Duley L et al (2008) Methods of prediction and prevention of pre-eclampsia: systematic reviews of accuracy and effectiveness literature with economic modelling. *Health Technol Assess* 12(6):1–270
3. Jacobi C, de Bock GH, Seigerink B, van Asperen CJ (2009) Differences and similarities in breast cancer risk assessment models in clinical practice: which model to choose? *Breast Cancer Res Treat* 115:381–390
4. Royston P, Moons KG, Altman DG, Vergouwe Y (2009) Prognosis and prognostic research: Developing a prognostic model. *Br Med J* 338:b604
5. Williams C, Brunskill S, Altman D, Briggs A, Campbell H, Clarke M et al. (2006) Cost-effectiveness of using prognostic information to select women with breast cancer for adjuvant systemic therapy. *Health Technol Assess* 10(34)
6. Altman D (2009) Prognostic models: a methodological framework and review of models for breast cancer. *Cancer Invest* 27:235–243
7. Chen WY, Colditz GA (2007) Risk factors and hormone-receptor status: epidemiology, risk-prediction models and treatment implications for breast cancer. *Nat Clin Pract Oncol* 4(7):415–423
8. Tice JA, Cummings SR, Smith-Bindman R, Ichikawa L, Barlow WE, Kerlikowske K et al (2008) Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model. *Ann Intern Med* 148(5):337–347
9. Rosner B, Colditz GA, Iglehart JD, Hankinson SE (2008) Risk prediction models with incomplete data with application to prediction of estrogen receptor-positive breast cancer: prospective data from the Nurses’ Health Study. *Breast Cancer Res* 10(4):R55
10. DerSimonian R, Laird N (1986) Meta-analysis in clinical trials. *Control Clin Trials* 7:177–188
11. Stata Corporation. Statistical software release 11.0. 2009. College Station, Texas
12. Riley RD, Higgins JP, Deeks JJ (2011) The interpretation of random effects meta-analysis. *Br Med J* 342:d549
13. Rockhill B, Byrne C, Rosner B, Louie MM, Colditz G (2003) Breast cancer risk prediction with a log-incidence model: evaluation of accuracy. *J Clin Epidemiol* 56(9):856–861

14. Viallon V, Ragusa S, Clavel-Chapelon F, Bénichou J (2009) How to evaluate the calibration of a disease risk prediction tool. *Stat Med* 28(6):901–916
15. Arne GN (2009) Breast cancer risk assessments to barrier contraception exposure. A new approach. *Makedonska Akademija na Naukite i Umetnostite Oddelenie Za Bioloshki i Meditsinski Nauki Prilozi* 30(1):217–232
16. Barlow WE, White E, Ballard-Barbash R, Vacek PM, Titus-Ernstoff L, Carney PA et al (2006) Prospective breast cancer risk prediction model for women undergoing screening mammography. *J Natl Cancer Inst* 98(17):1204–1214
17. Cook NR, Rosner BA, Hankinson SE, Colditz GA (2009) Mammographic screening and risk factors for breast cancer. *Am J Epidemiol* 170(11):1422–1432
18. Boyle PM (2004) Contribution of three components to individual cancer risk predicting breast cancer risk in Italy. *Eur J Cancer Prev* 13(3):183–191
19. Chen J, Pee D, Ayyagari R, Graubard B, Schairer C, Byrne C et al (2006) Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. *J Natl Cancer Inst* 98(17):1215–1226
20. Colditz GA, Rosner B (2000) Cumulative risk of breast cancer to age 70 years according to risk factor status: data from the Nurses Health Study. *Am J Epidemiol* 152(10):950–964
21. Rosner B, Colditz GA, Willett WC (1994) Reproductive risk factors in a prospective study of breast cancer: the Nurses' Health Study. *Am J Epidemiol* 139(8):819–835
22. Decarli A, Calza S, Masala G, Specchia C, Palli D, Gail MH et al (2006) Gail model for prediction of absolute risk of invasive breast cancer: independent evaluation in the Florence-European Prospective Investigation Into Cancer and Nutrition cohort. *J Natl Cancer Inst* 98(23):1686–1693
23. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C et al (1989) Projecting individualised probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* 81:1879–1886
24. Costantino JP, Gail MH, Pee D, Anderson S, Redmond CK, Benichou J et al (2010) Validation studies for models projecting the risk of invasive and total breast cancer incidence. *J Natl Cancer Inst* 91(18):1541–1548
25. Gail MH, Costantino JP, Pee D, Bondy M, Newmon L, Selvan M et al (2007) Projecting individualized absolute invasive breast cancer risk in African American women. *J Natl Cancer Inst* 99(23):1782–1792
26. Novotny J, Pecen L, Petruzalka L, Svobodnik A, Dusek L, Danes J et al (2006) Breast cancer risk assessment in the Czech female population—an adjustment of the original Gail model. *Breast Cancer Res Treat* 95:29–35
27. Rosner B, Colditz GA (1996) Nurses Health study: log-incidence mathematical model of breast cancer incidence. *J Natl Cancer Inst* 88(6):359–364
28. Tice JA, Cummings SR, Ziv E, Kerlikowske K (2005) Mammographic breast density and the Gail model for breast cancer risk prediction in a screening population. *Breast Cancer Res Treat* 94(2):115–122
29. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. [Erratum appears in *Stat Med*. 2005 Jan 15;24(1):156]. *Stat Med* 2004; 23(7): 1111–1130
30. Wacholder S, Hartge P, Prentice R, Garcia-Closas M, Feigelson HS, Diver WR et al (2010) Performance of common genetic variants in breast-cancer risk models. *N Engl J Med* 362(11):986–993
31. Bondy MLL (1994) Validation of a breast cancer risk assessment model in women with a positive family history. *J Natl Cancer Inst* 86(8):20
32. Costantino JP, Gail MH, Pee D, Anderson S, Redmond CK, Benichou J et al (1999) Validation of studies for models projecting the risk of invasive and total breast cancer incidence. *J Natl Cancer Inst* 91(18):1541–1548
33. Spiegelman DC (1994) Validation of the Gail et al. model for predicting individual breast cancer risk. *J Natl Cancer Inst* 86(8):20
34. Amir E, Evans DG, Shenton A, Lalloo F, Moran A, Boggis C et al (2003) Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme. *J Med Genet* 40(11):807–814
35. Rockhill B, Spiegelman D, Byrne C, Hunter DJ, Colditz GA (2001) Validation of the Gail et al. model of breast cancer risk prediction and implications for chemoprevention. *J Natl Cancer Inst* 93(5):358–366
36. Schonfeld SJ, Pee D, Greenlee RT, Hartge P, Lacy JV Jr, Park Y et al (2010) Effect of changing breast cancer incidence rates on the calibration of the Gail model. *J Clin Oncol* 28(14):2411–2417
37. Ulusoy C, Kepenekci I, Kose K, Aydinoglu S, Cam R (2010) Applicability of the gail model for breast cancer risk assessment in turkish female population and evaluation of breastfeeding as a risk factor. *Breast Cancer Res Treat* 120(2):419–424
38. Amir EFreedman (2010) Assessing women at high risk of breast cancer: A review of risk assessment models. *J Natl Cancer Inst* 102(10):680–691
39. Diez Collar MC, Ortega MP, Villanueva OR, Albaladejo VR, Astasio AP, Calle Puro ME et al (2000) Epidemiological characteristics of breast cancer development in pre and postmenopausal women [Spanish]. *Med Clin* 115(8):281–286
40. Mallett S, Royston P, Dutton S, Waters R, Altman D (2010) Reporting methods in studies developing prognostic models in cancer: a review. *BMC Med* 8:20
41. Mallett S, Royston P, Dutton S, Waters R, Altman D (2010) Reporting performance of prognostic models in cancer: a review. *BMC Med* 8:21
42. Janssens AC, Ioannidis J, van Duijn CM, Little J, Khoury MJ et al (2011) Strengthening the Reporting of Genetic Risk Prediction Studies: The GRIPS Statement. *PLoS Med* 8(3):e1000420
43. Collins G. Opening up multivariable prediction models. <http://blogs.bmj.com/bmj/2011/08/03/gary-collins-opening-up-multi-variable-prediction-models>. Accessed September 2011
44. Bleeker SE, Moll HA, Steyerberg EW, Donders AR, Derksen-Lubsen G, Grobbee DE et al (2003) External validation is necessary in prediction research: A clinical example. *J Clin Epidemiol* 56:826–832
45. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS et al (2008) Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. *PLoS Med* 5:e165
46. Agrawal R, Sharma S, Bekir J, Conway G, Bailey J, Balen AH et al (2004) Prevalence of polycystic ovaries and polycystic ovary syndrome in lesbian women compared with heterosexual women. *Fertil Steril* 82(5):1352–1357
47. Elmore JGF (2006) The risk of cancer risk prediction: “What is my risk of getting breast cancer?”. *J Natl Cancer Inst* 98(23):06
48. Cook NR (2010) Assessing the incremental role of novel and emerging risk factors. *Curr Cardiovasc Risk Rep* 4:112–119
49. Pencina MJ, D'Agostino RBS, D'Agostino RBJ, Vasan RS (2008) Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 27:157–172