Heliyon 9 (2023) e21344

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Review article

Review: Predictive approaches to breast cancer risk

Shuai Huang [a], Jun Tao Xu [b], Mei Yang [a,*]

[a] Department of Breast Oncology, Guangdong Provincial People's Hospital(Guangdong Academy of Medical Sciences), Southern Medical University, Guangdong, China

[b] Joint Turing-Darwin Laboratory of Phil Rivers Technology Ltd. and Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China Department of Computational Biology, Phil Rivers Technology Ltd, Beijing, China West Institute of Computing Technology, Chinese Academy of Sciences, Chongqing, China

ARTICLE INFO

Keywords:
Breast cancer
Risk factors
BRCA1/2
Polygenic risk scores
Artificial intelligence

ABSTRACT

Despite the deployment of specific breast cancer screening strategies, breast cancer incidence rates have escalated significantly over recent decades. In a bid to reverse this trend, scientists have engaged in extensive epidemiological research into breast cancer prevalence, identifying numerous individual risk factors and promoting population-wide health education. Coupled with advances in genetic testing, risk prediction models based on breast cancer genes have been developed, albeit with inherent limitations. In the new millennium, the emergence of artificial intelligence (AI) as a dominant technological force suggests that breast cancer prediction models developed with AI may represent the next frontier in research.

1. Introduction

As per a 2018 epidemiological report, breast cancer holds the unwelcome distinction of being the most prevalent cancer among women, contributing significantly to cancer-related mortality and disability-adjusted life years (DALYs) (1.7 million incident cases, 535,000 deaths, and 14.9 million DALYs) [1]. With the moniker of "the pink killer," breast cancer poses a grave threat to women's health globally. Efforts are underway to scientifically predict the risk of breast cancer in populations and guide individuals in formulating appropriate screening strategies. This precision prevention approach, tailored to varying risk profiles, optimizes the utilization of medical resources. Over the past several decades, endeavors to identify causative factors of cancer and prevent its onset have evolved. From initial epidemiological investigations of breast cancer risk factors to models combining genetics and risk, and now to contemporary AI-based breast cancer prediction models, each step underscores the relentless efforts to alleviate the disease burden of breast cancer. This article chronologically reviews these methodologies aimed at predicting the risk of breast cancer onset.

2. Non-hereditary individual risk factors

Given that the majority of breast cancer patients are women, the quest to understand its etiology has largely centered on attributes specific to females. Investigations that commenced in the 1950s studied reproductive factors such as the age of onset of menarche and menopause, pregnancy, and breastfeeding. The 1990s saw the emergence of additional risk factors including Body Mass Index (BMI), hormone replacement therapy, and pathological biopsy. Through these comprehensive studies, the scientific community aims to

* Corresponding author. Department of Breast Oncology, 4th Floor, 123 Huifu West Road, Yuexiu District, Guangzhou, Guangdong, China.
E-mail address: yangmei286@163.com (M. Yang).

https://doi.org/10.1016/j.heliyon.2023.e21344
Received 5 October 2023; Received in revised form 17 October 2023; Accepted 19 October 2023
Available online 31 October 2023

mitigate breast cancer risk by evading identified risk factors. The table below encapsulates several risk factors and their corresponding relative risks.

Risk Factor	Relative Risk
Age of first birth	Women who give birth to their first child under the age of 18 have only one-third the risk of breast cancer as those who delay childbirth until 35 or older [2]
Alcohol consumption	Daily drinking 5–9.9 g (equivalent to 3–6 drinks per week): 1.15, At least 30 g daily (at least 2 drinks daily): 1.51 [3]
Oral contraceptives	Current users: 1.24; 1–4 years after stopping: 1.16; 5–9 years after stopping: 1.07 [4]
Pregnancy and breastfeeding	For each childbirth: RR decreases by 7.0 %; For every 12 months of breastfeeding: RR decreases by 4.3 % [5]
Mammographic density	The higher the percentage density, the greater the risk of breast cancer : 5 %: 1 (baseline) ; 5 %–24 %: 1.79; 25%–49 %: 2.11; 50%–74 %: 2.92; Over 75 %: 4.64 [6]
Obesity	Pre-menopausal breast cancer risk is inversely related, RR is 0.54 (BMI>31 kg/m^2), for every increase of 5 kg/m^2 in BMI, the risk of pre-menopausal breast cancer decreases by about 8 %; Postmenopausal breast cancer risk is positively correlated with every increase of 5 kg/m^2 in BMI (RR, 1.12) [7]
Menarche and menopause	For every year menarche is advanced: RR increases by 1.050 times; For every year menopause is delayed: RR increases by 1.029 times [8]
benign breast disease	The summary risk estimate of developing breast cancer for non-proliferative disease was 1.17 (95 %CI 0.94–1.47). Proliferative disease without atypia was associated with significantly increased risk of future breast cancer, summary relative risk 1.76 (95 % CI 1.58–1.95). The summary risk estimate for atypical hyperplasia not otherwise specified was 3.93 (95 % CI 3.24–4.76) [9]
hormonereplacement therapy	Among menopausal hormone therapy users, these excess risks were definite even during years 1–4 (oestrogen-progestagen RR 1·60, 95 % CI 1·52–1·69; oestrogen-only RR 1·17, 1·10–1·26), and were twice as great during years 5–14 (oestrogen-progestagen RR 2·08, 2·02–2·15; oestrogen-only RR 1·33, 1·28–1·37) [10]

The factors related to breast cancer presented in the table are derived from epidemiological studies based on population exposure, and they have withstood the test of time and practical application. Most of these findings are analyzed through population-based epidemiology and aim to prevent breast cancer through public health education. They do not truly focus on individualized risk prevention strategies. Despite the increasing research into breast cancer risk factors, the incidence rate remains persistently high. Such risk assessment methods have not alleviated the socio-economic burden posed by breast cancer.

3. Advancement in developing breast cancer risk Prediction models

In the wake of profound research on breast cancer risk factors, a methodology has been devised to predict the risk of breast cancer, integratively taking into account relevant factors pinpointed in earlier studies. This method strives to estimate the probability of developing breast cancer for women at specific ages and with distinct risk factors, thereby facilitating bespoke screening and prevention plans for individuals and clinicians.

3.1. The GAIL model - a tool for assessing breast cancer risk

In 1989, Gail and his team introduced an innovative model [11]. The sample for this model was drawn from the Breast Cancer Detection Demonstration Project (BCDDP), encompassing 2,852 cases of white women and 3,146 white female controls. Within this dataset, a family history of breast cancer in first-degree relatives, age at first childbirth, age at menarche, and prior benign breast biopsies were identified as risk factors associated with breast cancer onset. Through specific mathematical transformations, coefficients for each risk factor were integrated, and the overall risk, *r(t),* for women was calculated using the formula

$$\mathrm{r}(t) = \prod_{i\in''factors''} r(t)_i \tag{1}$$

Subsequently, the baseline incidence rate $h_1(t)$ for breast cancer in women of different ages was determined from the BCDDP female data. The formula

$$h^*(t) = h_1(t)r(t) \tag{2}$$

was then applied to ascertain the incidence rate $h^*(t)$ for breast cancer in women of a specific age, considering the composite risk factor $r(t)$ in equation (1). The article also elaborated on how data from the GAIL model was used to estimate the probability of breast cancer occurrence in women of a specific age a over the next τ years, employing the integral formula

$$\mathrm{P}(a,\tau,r) = \int_a^{a+\tau} h^*(t)S(t)dt \tag{3}$$

Here, $S(t)$ represents the survival function, but considering other competing risk factors, $S(t)$ can be substituted with a more intricate survival function,and $h^{*}(t)$ comes from equation (2). Other factors such as HRT, drinking and smoking are related to breast cancer, but these factors are not sufficiently exposed in BCPT population, so the model does not consider other factors, and the selection of BCPT population is biased and cannot represent the general population.

In 1999, Gail and his team revised the model, substituting BCDDP data with SEER epidemiological data, thereby establishing a new baseline incidence rate $h_1(t)$ for women of varying ages [12]. The revised GAIL model (GAIL Model 2) was employed to predict the risk of invasive breast cancer and was validated over a five-year follow-up in 5,969 women from the Breast Cancer Prevention Trial (BCPT). The results indicated that GAIL Model 2 demonstrated commendable consistency in predicting invasive breast cancer and verified its absolute risk prediction accuracy over an average four-year follow-up period.

The GAIL model stands as the pioneering model with significant influence, derived from large-scale population data. However, its design did not incorporate the subsequently identified BRCA gene mutations. Consequently, for individuals known to carry BRCA mutations, the GAIL model might not offer precise predictions. Additionally, the model excludes women under the age of 35 and populations from non-Western countries, potentially leading to inaccuracies for women in this age bracket or those of other ethnicities. Nonetheless, given the GAIL model's foundation on easily measurable risk factors and its user-friendly nature, it continues to be employed clinically and has been endorsed by the NCCN guidelines as a preliminary tool for breast cancer risk assessment [13].

3.2. The BRCAPRO model - breast cancer risk assessment Program

The discovery of the BRCA1 gene in 1994 unveiled the familial genetic foundation of breast and ovarian cancers. For women with a family history of these cancers, the risk associated with carrying a BRCA1 gene mutation may be significantly elevated. To assist women in understanding their risk, Donald A. Berry and colleagues developed the BRCAPRO model in 1997 [14]. This model employs a combination of Mendelian genetics and Bayesian statistics to calculate the probability of a woman carrying the BRCA1 mutation based on her family history. The model's calculation is as follows:

Let M represent "the individual carries the mutation", N its complement "the individual does not carry the mutation", and H the individual's family history. The probability of an individual being a mutation carrier is P(M) as not considering the individual's family, where

$$P(M) = 2f - f^2 \tag{4}$$

Here, f is the allele frequency of the mutation in the evaluated individual's population, in the study population of this model, $f =$ 0.0006, thus

$$P(N) = 1 - P(M) \tag{5}$$

Applying Bayes' theorem and considering the individual's family history H,

$$P(M|H) = (P(H|M) \times P(M)) \ / \ P(H) \tag{6}$$

P(H) is the total probability of the individual's family history H, which can be decomposed as:

$$P(H) = P(H|M) \times P(M) + P(H|N) \times P(N) \tag{7}$$

Substituting the likelihood ratio,

$$LR = (P(H|M)) / \ P(H|N), \tag{8}$$

into the previous equations (6) and (7) yields the Bayesian formula

$$P(M|H) = LR \ / \ (LR + P(N) \ / \ P(M)) \tag{9}$$

Here, LR can be determined based on the epidemiological data of the individual's population, manifesting as different LR values for different ages. Combining this with Equation (4), (5) and (9) , we obtain the probability, $P(M|H)$,of an individual carrying a BRCA mutation given the family history H.

It's worth noting that when considering the exact relationships of all family members, the age of diagnosis for affected members, and the current age of unaffected members, the likelihood ratio LR undergoes formula adjustments. However, the core concept remains consistent, estimating the probability of specific family members carrying the BRCA1 mutation using estimated BRCA1 mutation frequencies in the general population and age-specific incidence rates for mutation carriers and non-carriers for both breast and ovarian cancers. While the initial version of the model did not account for BRCA2 mutations, subsequent versions were revised accordingly. Importantly, given that BRCA1/2 mutation-associated breast cancers constitute a minor fraction of all breast cancer cases, relying solely on BRCA1/2 does not fully explain the occurrence of hereditary breast cancers, limiting the predictive accuracy of the BRCAPRO model. A prospective prediction study published in 2019, encompassing four models, indicated that the BRCAPRO model's accuracy in predicting breast cancers in women without BRCA1/2 mutation inheritance was relatively low, with a ratio of expected to observed cases of 0.59 [95 % CI 0.55–0.64] [15].

3.3. OADICEA model - the breast and ovarian analysis of Disease incidence and carrier estimation algorithm

In the realm of familial breast cancer genetics, BRCA1 and BRCA2 stand out as the paramount "high-risk" genes. As research has advanced, mutations in the TP53, PTEN, and the relatively infrequent HRAS1 genes have been identified to be associated with the onset of breast cancer. Moreover, there might be other genes, still unidentified, that play a role in this context. To enhance the precision in predicting an individual's BRCA mutation status and their subsequent risk of developing breast cancer, Antoniou AC and colleagues, in 2002, devised the BOADICEA model [16]. This model bifurcates the risk estimation process into two segments.

Initially, it draws upon family history and employs the Bayesian approach to compute the probability, π_{j1}, of an individual harboring a BRCA mutation. The formula is given by

$$\pi_{j1} = \frac{P(\mathrm{BRCA1}, \mathrm{FH}_j)}{P(\mathrm{BRCA1}, \mathrm{FH}_j) + P(\mathrm{BRCA2}, \mathrm{FH}_j) + P(\mathrm{Non} - \mathrm{BRCA1}/2, \mathrm{FH}_j)} \tag{10}$$

Here, P(BRCA1, FH_j) represents the likelihood of observing a BRCA1 mutation in an individual given the family history FH_j. Similarly, P(BRCA2, FH_j) denotes the probability of observing a BRCA2 mutation in an individual under the same family history, while P(Non-BRCA1/2, FH_j) indicates the chance of observing a mutation that is neither BRCA1 nor BRCA2 under the family history FH_j. For detailed computations, the MENDEL software is recommended. By leveraging the likelihood ratio formula, the model calibrates the probability π_{j1} to best align with the BRCA testing outcomes from both the Anglian Breast Cancer (ABC) study cohort and the 156 high-incidence breast cancer families in the UK. This methodology facilitates the estimation of the likelihood that an individual might carry a BRCA mutation.

Another facet of the model postulates the existence of a putative BRCA3 gene, which, akin to BRCA1 and BRCA2 genes, plays a contributory role in the onset of breast and ovarian cancers. Researchers have predicated the incidence rates of breast and ovarian cancers on the Cox model, leading to the formulation of a mathematical equation:

$$\lambda(t) = \lambda_0(t) \times \exp(G + P) \tag{11}$$

This equation integrates age, genetic mutations, and other polygenic effects to estimate the incidence rates of breast and ovarian cancers. Here, the baseline incidence rate,$\lambda_0(t)$, is determined epidemiologically based on birth cohorts. The term 'G' represents the effects associated with the primary genotypes (BRCA1, BRCA2, and BRCA3). Meanwhile, *'P'* encapsulates the potential influences of numerous other genes on the risk of breast and ovarian cancers, amalgamating these effects into a single variable denoted as *'P'*. By maximizing the likelihood ratio formula, the model parameters are calibrated to align with observed data from both the Anglian Breast Cancer (ABC) study cohort and the 156 high-incidence breast cancer families. Furthermore, the BOADICEA model postulates age-specific relative risks of breast cancer for carriers of BRCA1 and BRCA2 mutations, offering a more precise estimation of the incidence rate curve. The model's predictions regarding the overall familial risk of breast cancer closely mirror epidemiological research findings, furnishing a sound basis for risk assessment in individuals with relatives diagnosed with breast or ovarian cancers.

Since 2005, Genome-Wide Association Studies (GWAS) have significantly advanced. GWAS is a technique used to identify genetic variations associated with diseases or traits by comparing genomic data from a large number of individuals. These variations can include single nucleotide polymorphisms (SNPs) or other types of variations like insertions/deletions [17]. By using GWAS, researchers can compare genomic data from numerous individuals to identify genetic variations linked to the incidence of disease. In 2007, Purcell et al. developed the Polygenic Risk Score (PRS) model based on these variations. PRS is a method that predicts individual disease risk by summing the effects of multiple genetic variations to calculate a risk score [18]. PRS has been widely applied in predicting various diseases, including in the field of cancer. For breast cancer prediction, genome-wide association studies analyze a large number of single nucleotide polymorphisms (SNPs) using high-throughput genotyping or sequencing technologies to detect their associations with breast cancer. By calculating a multi-gene risk score (PRS), researchers can predict an individual's risk of developing breast cancer and even anticipate the molecular subtypes of breast cancer [19].

In 2019, the BOADICEA model was updated to version 5.0.0 [20]. This iteration of the BOADICEA model incorporated the effects of the PALB2, CHEK2, and ATM genes, integrated a polygenic risk score (PRS) based on 313 single nucleotide polymorphisms, and included other non-genetic risk factors. The original 'G' effect, which was initially associated with BRCA1, BRCA2, and BRCA3, was expanded to encompass the genotypic effects of BRCA1, BRCA2, PALB2, CHEK2, and ATM. The 'P' variable was modified to combine the PRS with the effects of other yet unidentified genes, leading to a more refined stratification of breast cancer risk. A prospective validation study in 2021, involving 3,098 samples for the BOADICEA 5.0.0 version, revealed that the c-statistic (a diagnostic test evaluation metric) for BOADICEA improved from an initial 0.56 (95 % CI: 0.53–0.59) to 0.62 (95 % CI: 0.59–0.64) [21], indicating that the updated version offers enhanced guidance for preventive decision-making. The NICE guidelines have endorsed the BOADICEA model as a predictive tool for breast cancer risk [22]. Clinicians can access and utilize the model for patient evaluations via the dedicated website (www.canrisk.org), employing it as a clinical reference to guide individualized preventive strategies. However, despite its multiple revisions, the BOADICEA model has its limitations. Designed primarily based on European Caucasian populations, it may exhibit biases when applied to other ethnic groups, similar to the GAIL model. Additionally, the model does not account for all individual risk factors, notably lacking considerations for breastfeeding history and prior breast biopsy results.

3.4. IBIS model - international breast cancer intervention study

The initial design philosophy of the IBIS model stood in contrast to that of the GAIL model, which primarily focused on non-genetic factors of breast cancer, and the BRCAPRO model, which centered on genetic inheritance. Instead, the IBIS model combined individual risk factors with BRCA gene inheritance to assess the risk of breast cancer onset. In 2004, Jonathan Tyrer and colleagues developed a novel computational model termed the IBIS model [23]. This model proposed two primary genetic loci. The first locus is associated with the BRCA genes and could encompass the normal allele, the BRCA1 allele, or the BRCA2 allele. The second locus represents a hypothetical susceptibility gene, described as a "low-penetrance gene." This gene integrates the effects of several other genes, elevating the relative risk of breast cancer. Being dominant in nature, this low-penetrance gene ensures that females exhibit the same phenotype regardless of whether they carry one or two copies of the gene. Consequently, individuals can be mapped to one of the six phenotypes presented in Fig. 1(Fig. 1 depicts a table from Ref. [23]).

Following the assumption of the six phenotypes, Jonathan Tyrer and his team devised a three-step procedure to calculate an individual's risk of developing breast cancer:

1. Genotypic Probability P(xi|gi) Calculation: For each phenotype, the probability of individual developing breast cancer given genotype gi is calculated. The IBIS model employs Bayes' theorem to combine family history and population gene frequencies to compute individual genotype probabilities. This involves multiple probability calculations, including the probability of a phenotype given a genotype, the probability of a genotype given parental genotypes, and the genotype probability based on population gene frequencies. These probabilities are amalgamated using intricate mathematical methods to compute the likelihood for the entire pedigree, yielding individual genotype probabilities. The individual's gene probability P(gi) represents the probability that individual has genotype gi based on population gene frequencies. P(gi|gjgk) represents the probability that individual has genotype gi given parental genotypes gj and gk, calculated under the assumption that the genes are non-linked and autosomal.

The likelihood formula (12), is then combined with the individual's pedigree to calculate *P(x$_i$|g$_i$)*, representing the probability that individual i has a specific phenotype given genotype g_i.

$$L = \sum_{g_1} \cdots \sum_{g_m} \prod_{i=1}^{m} P(x_i|g_i)P(g_i|..) \tag{12}$$

2. Future Risk Fi(t$_1$;t$_2$) Calculation by Age and Genotype: The IBIS model employs a proportional hazard model to describe breast cancer development, using a survival function to calculate the risk of individuals with different genotypes developing breast cancer between ages t_i and t_j. The formula (13) is first used to determine *Snon(t)*, where *Snon(t)* and *Spop(t)* are the survival functions for females without the BRCA genes and the overall population, respectively. p_1 and p_2 represent the proportions of the population carrying the BRCA1 and BRCA2 mutations, respectively, with *Spop(t)*, p_1, and p_2 derived from UK national statistics. The baseline survival function *S$_0$(t)* is then calculated using

$$S_{\text{non}}(t) = (S_{\text{pop}}(t) - p_1 S_{\text{BRCA1}}(t) - p_2 S_{\text{BRCA2}}(t))/(1 - p_1 - p_2) \tag{13}$$

$$\sum_q p_g S_0(t)^{\theta_g} = S_{\text{non}}(t), \tag{14}$$

where p_g represents the population proportion with genotype *g*. Once S_0(t) is determined using Newton–Raphson method, the risk for an individual with a particular genotype to develop breast cancer between ages t_i and t_j is calculated as $S_0(t_i)^{\theta_g} - S_0(t_j)^{\theta_g}$. Consequently, for those without any risk gene copies, the risk is $(S_0(t_i){-}S_0(t_j))$. For individuals carrying at least one copy of the risk gene, the risk is $(S_0(t_i)\theta{-}S_0(t_j)\theta)$. For carriers of the BRCA1 gene, the risk of developing breast cancer between ages t_i and t_j for a specific genotype is $(S_1(t_i)^{\theta_g} - S_1(t_j)^{\theta_g})$. Similarly, for BRCA2 carriers, the risk is $(S_2(t_i)^{\theta_g} - S_2(t_j)^{\theta_g})$. Using UK national statistics, Jonathan Tyrer and colleagues determined gene frequencies of 0.11 and 0.12% for BRCA1 and BRCA2, respectively. By fitting the model to population data from Anderson et al. (population data from reference (24)), they inferred that approximately 21 % of the population carries the hypothesized low-penetrance gene, with a relative risk θ of 13.0377 for those carrying at least one low-penetrance gene.

3. **Calculating the Composite Risk α by Integrating Individual Risk Factors:** Jonathan Tyrer and colleagues evaluated the relative risks associated with individual risk factors such as age at menarche, age at menopause, and parity. An individual's overall risk

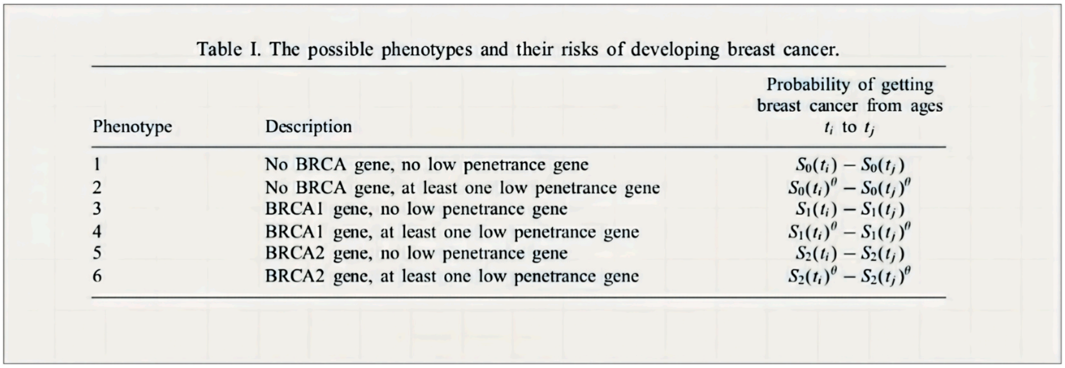

Table I. The possible phenotypes and their risks of developing breast cancer.

Phenotype	Description	Probability of getting breast cancer from ages t_i to t_j
1	No BRCA gene, no low penetrance gene	$S_0(t_i) - S_0(t_j)$
2	No BRCA gene, at least one low penetrance gene	$S_0(t_i)^\theta - S_0(t_j)^\theta$
3	BRCA1 gene, no low penetrance gene	$S_1(t_i) - S_1(t_j)$
4	BRCA1 gene, at least one low penetrance gene	$S_1(t_i)^\theta - S_1(t_j)^\theta$
5	BRCA2 gene, no low penetrance gene	$S_2(t_i) - S_2(t_j)$
6	BRCA2 gene, at least one low penetrance gene	$S_2(t_i)^\theta - S_2(t_j)^\theta$

Fig. 1. TableI from Reference [23].

α is the product of the ratio of each risk *f* for an individual to the corresponding risk *R* in the general population. This can be determined from specific epidemiological data.

To conclude, the formula:

$$\Pr(\ cancer\ ) = 1 - \left(1 - \sum_{i=1}^{6} p_i F_i(t_1, t_2)\right)^{\alpha} \tag{15}$$

is employed to calculate an individual's risk of developing breast cancer in the next 10 years or the cumulative risk up to 85 years of age. Here, p_i denotes the probability of phenotype i, equivalent to the probability $P(x_i|g_i)$ of exhibiting phenotype i given genotype g_i calculated in equation (14). $Fi(t_1;t_2)$ represents the probability derived in the second step, indicating the risk of contracting breast cancer between ages t_1 and t_2 given the woman's phenotype i. Lastly, α is the result calculated in the third step, representing the composite risk based on individual risk factors. Individuals are stratified based on the resulting probabilities $\Pr(\ cancer\ )$, and appropriate preventive plans are formulated according to their risk levels.

In 2018, the IBIS model was updated to version 8.0 [25]. This version incorporated 18 single nucleotide polymorphisms (SNPs) associated with breast cancer and mammographic density into the model. A female population of 9,363 was used for validation. The specific method involved calculating the multi-gene risk score (SNP18) and breast density for each participant and performing calibration. Then, the 10-year breast cancer risk probability was calculated using the IBIS (version 6.0) model based on risk factors. Finally, the IBIS (version 6.0) 10-year absolute risk was multiplied by the calibrated breast density and PRS (SNP18) to obtain the combined 10-year risk probability. This process adjusted the risk stratification of the original 10-year risk components from IBIS (version 6.0), resulting in the reclassification of some patients from high risk to low risk and vice versa. As a result, the latest version of IBIS provides better risk stratification for predicting breast cancer in the population. In a prospective prediction study comparing four models published in 2019, the IBIS model performed well, with a ratio of expected cases to observed cases of 1.03 [95 % CI 0.96–1.12] [15].

The current NCCN guidelines recommend the IBIS model for predicting breast cancer risk [13]. By accessing the website provided by the NCCN guidelines and completing a questionnaire, individuals can assess their risk of developing breast cancer. The NCCN uses a 20 % risk threshold predicted by the IBIS model to categorize individuals into high and low-risk groups, thereby formulating different preventive strategies. However, the NCCN does not advocate for the integration of PRS with the IBIS model in clinical applications [13], given that PRS has not yet been successfully validated in clinical trials. The IBIS model, by incorporating a wide range of factors, offers a more comprehensive prediction of breast cancer risk. It not only predicts an individual's risk but also estimates the risk for first, second, and third-degree relatives (cousins). As a result, when compared to models like the Gail model, the IBIS model demonstrates superior performance in large-scale population validations. However, there are some challenges in its clinical application. For instance, the IBIS model cannot predict the risk for individuals who have previously undergone chest radiation therapy. Additionally, it tends to overestimate the risk for populations with atypical hyperplasia or high breast density.

3.5. BCSC Model - Breast Cancer surveillance consortium

In 2005, it was conclusively established that breast density is an independent risk factor for breast cancer [6], a factor not considered by previous models. Recognizing this oversight, Jeffrey A. Tice and colleagues introduced and validated an efficient, pragmatic breast cancer risk prediction model in 2008, that duly incorporates breast density [26]. The model merges factors such as age, race or ethnicity, family history of breast cancer, and breast biopsy history to compute the associated risk. The model categorizes the population into four groups based on breast density: almost entirely fatty (Category 1), scattered fibroglandular density (Category 2), heterogeneously dense (Category 3), and extremely dense (Category 4). Risk prediction of breast cancer is subsequently conducted for these four density categories. During a follow-up study conducted over 5.3 years, the model demonstrated the ability to distinguish prospectively between individuals who would develop breast cancer and those who would not, within a sample of 1,095,484 women. The breast density model showcased commendable calibration both in the overall population (with an expected-to-observed ratio of 1.03 [95 % CI, 0.99–1.06]) and across racial and ethnic subgroups. It also exhibited moderate discriminatory accuracy, with a concordance index of 0.66 [CI, 0.65–0.67].

Furthermore, several other models have been studied, but they have not been designed and validated using large-scale samples like the GAIL model, BRCAPRO model, BOADICEA model, and IBIS model, which have broader applications. It is worth mentioning that the BOADICEA model and IBIS model incorporate Polygenic Risk Scores (PRS) to enhance the predictive ability of the models. However, PRS also has some limitations. The calculation of PRS usually involves two steps. The first step is selecting a set of SNPs (single nucleotide polymorphisms) associated with a specific phenotype, which is typically identified through genome-wide association studies (GWAS) or other genetic studies. The second step is assigning a weight (often β coefficients) to each SNP to calculate an overall score. This score is obtained by multiplying the genotype (usually coded as 0, 1, or 2 representing the number of mutant alleles) of each SNP by its corresponding weight and summing the results. A higher overall score indicates a higher risk of the phenotype. However, different studies may use different sets of SNPs and values for β coefficients to calculate PRS, so the specific calculation methods may vary [27]. Additionally, PRS is influenced by the population of the samples. Currently, GWAS studies are mainly conducted in European and American populations, and the β coefficients of the same SNP may differ in Asian populations. The accuracy of PRS depends on whether the SNPs used to calculate it are applicable to the target population. Using SNPs from populations of different ethnicities may result in bias in the calculated PRS [28].

4. Advancements in the application of artificial intelligence for breast cancer prediction in the contemporary era

As we enter a new era, the accelerated development of artificial intelligence (AI) has paved the way for its integration into the prediction of breast cancer risk. Implementing AI techniques can refine the accuracy in assessing a patient's breast cancer risk and further facilitate personalized preventative measures.

4.1. TRS model - thermalytix risk score

In 2019, researchers from India, aiming to address the challenges of breast cancer screening in developing countries, introduced the Thermalytix Risk Score framework, which combines metabolic imaging omics with artificial intelligence [29]. This framework is divided into two parts:

The first part employs a Random Forest (RF) classifier based on 31 features, such as irregular shapes, elevated temperatures, and asymmetry, to identify individual metabolic abnormalities, resulting in the calculation of a "hotspot score." The second part uses a complex formula to reconstruct breast vascular images. Another Random Forest (RF) classifier is then applied to extract abnormal vessels from these images, leading to the computation of a "vascular score."

The framework utilizes deep learning to assign scores to patients based on abnormalities in breast tissue metabolism and vascular anomalies (these abnormal images appear before the formation of solid breast tumors). These scores are then combined, weighted differently, and a linear formula is used to compute the final Thermalytix Risk Score, which predicts an individual's risk of developing breast cancer. Based on the TRS, the 769 study participants were divided into four risk quartiles of equal width [0–0.25: Low Risk, 0.25–0.5: Medium Risk, 0.5–0.75: High Risk, 0.75–1.0: Very High Risk]. This categorization effectively predicted the risk of breast cancer in the population (with an Area Under the Curve, AUC, of 0.895 for the TRS), thereby facilitating better individualized screening plans and optimizing the use of medical resources.

The TRS model, which relies on less complex input data compared to other models like the GAIL and IBIS, holds the potential to significantly reduce costs associated with population-wide screening. This model is especially beneficial for use in regions with constrained medical resources, offering preventive consultation for breast cancer. However, the TRS model's viability for generating 5-year or 10-year risk predictions in larger populations warrants further exploration and confirmation.

4.2. DL model - deep learning mammography-based model

Breast density is a well-known risk factor for breast cancer, yet the subjective assessment of breast radiographic density lacks uniformity across radiology practices. Furthermore, there may be subtle yet informative indications present in mammograms that might elude human detection or simplistic volumetric density measurements. In 2019, Adam Yala et al. developed a deep learning model, the Deep Learning Mammography-based Model, which employs a deep convolutional neural network based on PyTorch to predict breast cancer risk using mammography images from 31,806 women [30]. The model showcased robust performance in predicting breast cancer risk in a validation population of 3,804 women and a testing population of 3,978 women. Surpassing the IBIS model (AUC = 0.62), the DL model achieved an AUC of 0.68. Furthermore, the DL model can be integrated with traditional risk factors such as age at menarche and parity to augment the model's AUC to 0.70. The DL model provides a convenient and cost-effective tool for screening patients, especially those with limited access to additional information. Nevertheless, there is still limited understanding regarding the precise mammographic information employed by the DL model in predicting risk, thus warranting further investigation to elucidate the underlying mechanisms. Moreover, it is imperative to conduct future studies to assess the performance of the DL model when integrated with genetic testing for breast cancer risk prediction.

4.3. Damage assessment of genomic mutations (DAGM) framework

In 2021, a novel approach was taken to investigate the relationship between germline genomic mutations and the risk of breast cancer using AI. The Damage Assessment of Genomic Mutations (DAGM) framework was established, employing AI technology to create a computational model [31]. The DAGM model inputs all germline mutations of an individual and calculates their cumulative effects on gene expression as reflected in cellular signaling pathways. Based on this information, the framework generates a Activity Profiles of Signaling Pathways for each individual and calculates an APSP (Activity Profiles of Signaling Pathways) score. The APSP score, indicating the impact of germline mutations on an individual's cellular signaling pathways, can be used to evaluate their risk of breast cancer. Although the DAGM model shows potential, it heavily relies on a framework constructed using publicly available data from the Catalogue of Somatic Mutations in Cancer (COSMIC) and has not been extensively validated in real-world populations. Moreover, it does not account for other individual risk factors in the database, and further improvements may necessitate adjustments to the model's calculation formula based on these additional factors.

Presently, there are numerous AI-based models under development. While most are still in their infancy and require further prospective validation studies for clinical application, the rapid evolution of AI technology undeniably forecasts an increase in model research. Consequently, the future of AI models in breast cancer risk prediction is promising. These models are anticipated to provide convenient and accurate assistance in predicting cancer risks, enabling individuals to make informed decisions.

5. Conclusion

The strategies for breast cancer prevention have undergone a transformation, evolving from preliminary epidemiological studies to sophisticated statistical methodologies, and now to contemporary AI-based predictive models. This transition not only mirrors the advancements in societal technology but also illustrates the continuous endeavors of humanity in disease prevention and treatment.

Given the significant benefits of early treatment for breast cancer patients, early diagnosis and monitoring have become crucial components. The current predictive models offer invaluable screening and examination recommendations to clinicians and high-risk individuals, empowering them to make informed decisions. These models enhance the quality of life for patients while optimizing the utilization of public health resources. By integrating various factors such as demographics, biomarkers, and genetics, these models provide more precise monitoring suggestions for high-risk individuals to physicians. Targeted screening and monitoring enable the early detection of potential diseases, offering more effective treatment options. At present, predictive systems like the GAIL and IBIS models have been widely adopted in clinical settings. Although these methods have achieved some success in reducing the incidence of breast cancer, there is still room for improvement in their detection capabilities. Emerging AI models have demonstrated exceptional predictive performance; however, their AUC may vary in broader populations, necessitating further clinical studies to validate their accuracy.

In conclusion, early diagnosis and treatment of breast cancer are pivotal in improving the survival rates and quality of life for patients. The current preventive measures and predictive models provide solid support in this regard. Through these means, we can more effectively identify high-risk individuals and offer them timely diagnosis and treatment, which is crucial for the control and prevention of breast cancer.

Data availability statement

This article is a review and did not generate or analyze any new datasets. The responses in the submission questionnaire regarding data availability were provided based on the nature of this review. Any data referenced in this review can be obtained from the original published sources.

CRediT authorship contribution statement

Shuai Huang: Writing – original draft, Formal analysis, Data curation. **Jun Tao Xu:** Formal analysis, Data curation. **Mei Yang:** Writing – review & editing, Writing – original draft, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We sincerely acknowledge the financial support provided for this research by the following institutions:

National Natural Science Foundation of China (Grant Nos. 82072939, 82373427).

Natural Science Foundation of Guangdong Province (Grant No. 2022A1515010425).

Guangzhou Science and Technology Program (Grant No. 202206010110).

We express our deep gratitude for the generous support from the aforementioned institutions.

References

[1] Global Burden of Disease Cancer Collaboration, et al., Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2016: a systematic analysis for the Global Burden of Disease study, JAMA Oncol. 4 (2018) 1553–1568.

[2] B. MacMahon, P. Cole, T.M. Lin, et al., Age at first birth and breast cancer risk, Bull. World Health Organ. 43 (2) (1970) 209–221. PMID: 5312521; PMCID: PMC2427645.

[3] W.Y. Chen, B. Rosner, S.E. Hankinson, G.A. Colditz, W.C. Willett, Moderate alcohol consumption during adult life, drinking patterns, and breast cancer risk, JAMA 306 (17) (2011) 1884–1890, https://doi.org/10.1001/jama.2011.1590. PMID: 22045766; PMCID: PMC329234.

[4] Collaborative Group on Hormonal Factors in Breast Cancer, Breast cancer and hormonal contraceptives: collaborative reanalysis of individual data on 53 297 women with breast cancer and 100 239 women without breast cancer from 54 epidemiological studies, Lancet 347 (9017) (1996) 1713–1727, https://doi.org/10.1016/s0140-6736(96)90806-5. PMID: 8656904.

[5] Collaborative Group on Hormonal Factors in Breast Cancer, Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and 96973 women without the disease, Lancet 360 (9328) (2002) 187–195, https://doi.org/10.1016/S0140-6736(02)09454-0. PMID: 12133652.

[6] N.F. Boyd, J.M. Rommens, K. Vogt, V. Lee, J.L. Hopper, M.J. Yaffe, A.D. Paterson, Mammographic breast density as an intermediate phenotype for breast cancer, Lancet Oncol. 6 (10) (2005) 798–808, https://doi.org/10.1016/S1470-2045(05)70390-9. Erratum in: Lancet Oncol. 2005 Nov;6(11):826. PMID: 16198986.

[7] M. Picon-Ruiz, C. Morata-Tarifa, J.J. Valle-Goffin, E.R. Friedman, J.M. Slingerland, Obesity and adverse breast cancer risk and outcome: mechanistic insights and strategies for intervention, CA Cancer J Clin 67 (5) (2017) 378–397, https://doi.org/10.3322/caac.21405. Epub 2017 Aug 1. PMID: 28763097; PMCID: PMC5591063.

[8] Collaborative Group on Hormonal Factors in Breast Cancer, Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies, Lancet Oncol. 13 (11) (2012) 1141–1151, https://doi.org/10.1016/S1470-2045(12)70425-4. Epub 2012 Oct 17. PMID: 23084519; PMCID: PMC3488186.

[9] S.W. Dyrstad, Y. Yan, A.M. Fowler, G.A. Colditz, Breast cancer risk associated with benign breast disease: systematic review and meta-analysis, Breast Cancer Res. Treat. 149 (3) (2015) 569–575, https://doi.org/10.1007/s10549-014-3254-6.

[10] Type and timing of menopausal hormone therapy and breast cancer risk: individual participant meta-analysis of the worldwide epidemiological evidence, Lancet (2019), https://doi.org/10.1016/s0140-6736(19)31709-x.

[11] M.H. Gail, L.A. Brinton, D.P. Byar, D.K. Corle, S.B. Green, C. Schairer, J.J. Mulvihill, Projecting individualized probabilities of developing breast cancer for white females who are being examined annually, J. Natl. Cancer Inst. 81 (24) (1989) 1879–1886, https://doi.org/10.1093/jnci/81.24.1879. PMID: 2593165.

[12] J.P. Costantino, M.H. Gail, D. Pee, S. Anderson, C.K. Redmond, J. Benichou, H.S. Wieand, Validation studies for models projecting the risk of invasive and total breast cancer incidence, J. Natl. Cancer Inst. 91 (18) (1999) 1541–1548, https://doi.org/10.1093/jnci/91.18.1541. PMID: 10491430.

[13] NCCN Clinical Practice Guidelines in Oncology: Breast Cancer Risk Reduction. Available at: [https://www.nccn.org/professionals/physician_gls/pdf/breast_risk.pdf]. Accessed on [2023/11/02], page 19.

[14] D.A. Berry, G. Parmigiani, J. Sanchez, J. Schildkraut, E. Winer, Probability of carrying a mutation of breast-ovarian cancer gene BRCA1 based on family history, J. Natl. Cancer Inst. 89 (3) (1997) 227–238, https://doi.org/10.1093/jnci/89.3.227. PMID: 9017003.

[15] Terry MB, Liao Y, et al. 10-year performance of four models of breast cancer risk: a validation study. Lancet Oncol..;20(4): 504-517. doi:10.1016/S1470-2045(18)30902-1. Epub 2019 Feb 21. PMID: 30799262..

[16] A.C. Antoniou, P.D. Pharoah, G. McMullan, N.E. Day, M.R. Stratton, J. Peto, B.J. Ponder, D.F. Easton, A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes, Br. J. Cancer 86 (1) (2002) 76–83, https://doi.org/10.1038/sj.bjc.6600008. PMID: 11857015; PMCID: PMC2746531.

[17] J. Hardy, A. Singleton, Genomewide association studies and human disease, N. Engl. J. Med. 360 (17) (2009) 1759–1768, https://doi.org/10.1056/NEJMra0808700. Epub 2009 Apr 15. PMID:19369657; PMCID: PMC3422859.

[18] S.M. Purcell, N.R. Wray, J.L. Stone, P.M. Visscher, M.C. O'Donovan, P.F. Sullivan, P. Sklar, Common polygenic variation contributes to risk of schizophrenia and bipolar disorder, Nature 460 (7256) (2009) 748–752.

[19] N. Mavaddat, K. Michailidou, J. Dennis, et al., Polygenic risk scores for prediction of breast cancer and breast cancer subtypes, Am. J. Hum. Genet. 104 (1) (2019) 21–34, https://doi.org/10.1016/j.ajhg.2018.11.002. Epub 2018 Dec 13. PMID: 30554720; PMCID: PMC6323553.

[20] A. Lee, N. Mavaddat, A.N. Wilcox, et al., BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors, Genet. Med. 21 (8) (2019) 1708–1718, https://doi.org/10.1038/s41436-018-0406-9. Epub 2019 Jan 15. Erratum in: Genet Med. 2019 Feb 21;: PMID: 30643217; PMCID: PMC6687499.

[21] S.X. Li, R.L. Milne, T. Nguyen-Dumont, et al., Prospective evaluation of the addition of polygenic risk scores to breast cancer risk models, JNCI Cancer Spectr. 5 (3) (2021) pkab021, https://doi.org/10.1093/jncics/pkab021. PMID: 33977228; PMCID: PMC8099999.

[22] National Institute for Health and Care Excellence (NICE), Familial Breast Cancer: Classification, Care and Managing Breast Cancer and Related Risks in People with a Family History of Breast Cancer (CG164), 2019 [PDF]. Last updated: 20 November 2019. Retrieved from [https://www.nice.org.uk/guidance/cg164.

[23] J. Tyrer, S.W. Duffy, J. Cuzick, A breast cancer prediction model incorporating familial and personal risk factors, Stat. Med. 23 (7) (2004) 1111–1130, https://doi.org/10.1002/sim.1668. Erratum in: 12 Stat Med. 2005 Jan 15;24(1):156. PMID: 15057881.

[24] H. Anderson, A. Bladström, H. Olsson, T.R. Möller, Familial breast and ovarian cancer: a Swedish population-based register study, Am. J. Epidemiol. 152 (12) (2000) 1154–1163, https://doi.org/10.1093/aje/152.12.1154.

[25] E.M. van Veen, A.R. Brentnall, H. Byers, E.F. Harkness, S.M. Astley, S. Sampson, A. Howell, W.G. Newman, J. Cuzick, D.G.R. Evans, Use of single-nucleotide polymorphisms and mammographic density plus classic risk factors for breast cancer risk prediction, JAMA Oncol. 4 (4) (2018) 476–482, https://doi.org/10.1001/jamaoncol.2017.4881. PMID: 29346471; PMCID: PMC5885189.

[26] J.A. Tice, S.R. Cummings, R. Smith-Bindman, L. Ichikawa, W.E. Barlow, K. Kerlikowske, 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) (2008) 337–347, https://doi.org/10.7326/0003-4819-148-5-200803040-00004. PMID: 18316752; PMCID: PMC2674327.

[27] J.C. Denny, L. Bastarache, M.D. Ritchie, R.J. Carroll, R. Zink, J.D. Mosley, S.J. Hebbring, Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data, Nat. Biotechnol. 31 (12) (2013) 1102–1110.

[28] A.V. Khera, M. Chaffin, K.G. Aragam, M.E. Haas, C. Roselli, S.H. Choi, P. Natarajan, E.S. Lander, S.A. Lubitz, P.T. Ellinor, S. Kathiresan, Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations, Nat. Genet. 50 (9) (2018) 1219–1224, https://doi.org/10.1038/s41588-018-0183-z. Epub 2018 Aug 13. PMID: 30104762; PMCID: PMC6128408.

[29] S.T. Kakileti, H.J. Madhu, G. Manjunath, L. Wee, A. Dekker, S. Sampangi, Personalized risk prediction for breast cancer pre-screening using artificial intelligence and thermal radiomics, Artif. Intell. Med. 105 (2020), 101854, https://doi.org/10.1016/j.artmed.2020.101854. Epub 2020 Apr 7. PMID: 32505418.

[30] A. Yala, C. Lehman, T. Schuster, T. Portnoi, R. Barzilay, A deep learning mammography-based model for improved breast cancer risk prediction, Radiology 292 (1) (2019) 60–66, https://doi.org/10.1148/radiol.2019182716. Epub 2019 May 7. PMID: 31063083.

[31] M. Yang, Y. Fan, Z.Y. Wu, J. Gu, Z. Feng, et al., DAGM: a novel modelling framework to assess the risk of HER2-negative breast cancer based on germline rare coding mutations, EBioMedicine 69 (2021), 103446, 10.1016/j.ebiom.2021.103446. Epub 2021 Jun 19. PMID: 34157485; PMCID: PMC8220579.