Risk Models for Breast Cancer and Their Validation

Adam R. Brentnall and Jack Cuzick

Abstract. Strategies to prevent cancer and diagnose it early when it is most treatable are needed to reduce the public health burden from rising disease incidence. Risk assessment is playing an increasingly important role in targeting individuals in need of such interventions. For breast cancer many individual risk factors have been well understood for a long time, but the development of a fully comprehensive risk model has not been straightforward, in part because there have been limited data where joint effects of an extensive set of risk factors may be estimated with precision. In this article we first review the approach taken to develop the IBIS (Tyrer–Cuzick) model, and describe recent updates. We then review and develop methods to assess calibration of models such as this one, where the risk of disease allowing for competing mortality over a long follow-up time or lifetime is estimated. The breast cancer risk model model and calibration assessment methods are demonstrated using a cohort of 132,139 women attending mammography screening in the State of Washington, USA.

Key words and phrases: Breast cancer, calibration, risk assessment, breast density, Tyrer–Cuzick model, IBIS model.

1. INTRODUCTION

Unlike lung cancer (tobacco smoking) and cervix cancer (persistent infection with the human papilloma virus) where a single factor explains the majority of the cases, a large number of factors have been found to be important for determining the risk of breast cancer. An increased risk in women with a family history appears to have been known in ancient Roman times [59], and in 1842 Rigoni-Stern reported that nuns had an increased risk of breast cancer [53, 26], paving the way for further research which established that not having a child or having a first childbirth at an older age increased the risk of this disease. Much of the modern work on risk factors was done by Macmahon and colleagues, mostly at Harvard [43]. Pike and colleagues emphasized the link with oestrogens and proposed that much of the population risk could explained by computing the cumulative lifetime exposure to oestrogens based mostly on age at menarche, childbearing

history and age at menopause [49]; an important meta-analysis and update on these and other factors was provided by Beral and colleagues [47].

In 1976, Wolfe discovered that breast density was another key risk factor that was broadly unrelated to the known classical factors, but of roughly equal importance in their combined predictive value [69]. This has been further developed by Boyd [11, 10], McCormack and others [45]. In the same paper, Wolfe also suggested [69] using breast density for risk assessment to determine how best to use mammography for breast cancer screening. Several decades later, advances in the ability to accurately stratify women into higher and lower-risk groups are likely to move early detection strategies towards this vision, and replace 'one-size-fits-all' screening with so-called 'risk-adapted' programs in which both the frequency with which a woman is screened and the modality are chosen based on the risk of breast cancer. Motivations include the identification of women at extremely high risk, who are potential candidates for risk-reducing surgery or preventive therapy [29], delineation of a group at moderately enhanced risk who might benefit from enhanced screening [60], and identification of populations at sufficiently low risk so as to require less frequency or even no screening.

A number of risk models have been developed, mostly for Caucasian women living in North America and Western Europe [32, 18, 4]. In this article, we focus on risk

Adam R. Brentnall is Lecturer in Biostatistics, Centre for Cancer Prevention, Wolfson Institute of Preventive Medicine, Queen Mary University of London, Charterhouse square, London, EC1M 6BQ (e-mail: a.brentnall@qmul.ac.uk). Jack Cuzick is John Snow Professor of Epidemiology, Centre for Cancer Prevention, Wolfson Institute of Preventive Medicine, Queen Mary University of London, Charterhouse square, London, EC1M 6BQ (e-mail: j.cuzick@qmul.ac.uk).

assessment using the Tyrer–Cuzick model [63]. This is a hybrid of two popular sub-models often used for breast cancer risk assessment: a genetic segregation model for familial risk that is combined with a proportional-hazards regression model for other risk factors.

Breast cancer risk factors in the model broadly fit into five general categories: (1) family history and highly penetrant dominant genetic mutations; (2) factors associated with oestrogen exposure, including age at first childbirth, age at menopause and menarche (beginning of periods), use of hormone replacement therapy, height and weight; (3) certain types of prior benign breast disease, (4) breast imaging features seen on the mammogram—notably the amount of dense tissue (opaque areas on a mammogram or breast x-ray); and (5) common but individually less penetrant genetic differences (single nucleotide polymorphisms, SNPs), where several hundred relatively common genetic variants that individually have a small impact on disease have been identified, and that jointly make an important contribution to overall risk assessment via a 'polygenic risk score' [44]. There are also some other apparent risk factors which are harder to quantify, but which may improve the performance of the model.

In the rest of this paper, we outline the statistical basis of our model, review and develop methods to assess model calibration that include competing mortality (which has been handled differently in different papers), and apply our model and calibration assessment methods to data from a large cohort of women attending mammographic screening in the State of Washington, USA.

2. BREAST CANCER RISK ASSESSMENT

2.1 Background

To provide an assessment of the risk of disease occurring within a person's residual lifetime it is important to take into account competing risks that could lead to death from other causes. In a classical 'latent lifetimes' framework [20] each individual is subject to m potential causes of death with times T_j ($j = 1, \ldots, m$) and we observe their actual time of death $\tilde{T} = \min(T_1, \ldots, T_m)$ and the specific cause $J = 1, \ldots, m$. This framework assumes that each death is attributable to a single cause, or a defined group of causes. We extend this approach to also include the incidence of breast cancer, which is our principle interest, and do not consider death after breast cancer occurs.

Multiple modes of failure are conveniently characterised using functions for the rate at which each cause J occurs at each follow-up time, given that the person has not yet died, or experienced a specific event of interest. More precisely, this *cause-specific hazard* for $j = 1, \ldots, m$ is defined as

$$(1) \qquad h_j(t) = \lim_{\Delta \to 0} \frac{\mathrm{P}(T_j < t + \Delta \mid \tilde{T} \geq t)}{\Delta}.$$

Equation (1) is estimable due to conditioning on $(\tilde{T} \geq t)$. Another measure of risk that is sometimes considered is the proportion of the population who have event J when followed up to time t. This *cumulative incidence function* is defined by

$$(2) \qquad \mathrm{P}(J = j, \tilde{T} \leq t) = \int_0^t h_j(u) S(u) \, \mathrm{d}u,$$

where

$$S(t) = \mathrm{P}(\tilde{T} > t)$$
$$= \exp\left\{ -\int_0^t \sum_{k=1}^m h_k(u) \, \mathrm{d}u \right\}$$

is the marginal (all-cause) survivor function for $\tilde{T}$. When there is just one cause (death, say) then (2) is the cumulative distribution function (for time of death).

Breast cancer risk models within this framework typically consider two causes ($m = 2$): T_1, the time to diagnosis of breast cancer and T_2, the time to death from other causes (excluding breast cancer mortality, which can't occur before the time of diagnosis), so that $\tilde{T} = \min(T_1, T_2)$. Risk assessment for a women usually takes place when current age $t_0 > 20$ years, and the aim is to assess the absolute risk of breast cancer between age t_0 to age t, conditional on q risk factors $\boldsymbol{x} = (x_1, \ldots, x_q)$. This may be done by extending (2) to condition on risk factors $\boldsymbol{x}$ and age at risk assessment t_0 through

$$
\begin{aligned}
&\mathrm{P}(J = 1, \tilde{T} \leq t \mid \boldsymbol{x}, \tilde{T} > t_0) \\
&= \int_{t_0}^t h_1(u \mid \boldsymbol{x}) \exp\left[-\int_{t_0}^u \{ h_1(v \mid \boldsymbol{x}) \right. \\
&\qquad \left. + h_2(v \mid \boldsymbol{x}) \} \, \mathrm{d}v \right] \mathrm{d}u \\
&\equiv P_x(t_0, t),
\end{aligned}
$$

(3)

where $h_1(t \mid \boldsymbol{x})$ is the conditional hazard of breast cancer at age t (cf. equation 1) and $h_2(t \mid \boldsymbol{x})$ is the conditional hazard for competing mortality. In practice, and following [32], competing mortality has often been taken to depend only on age and has been calibrated using national mortality statistics excluding breast cancer. This is reasonable for breast cancer, as it is not strongly linked other causes of death, but is more problematic, for example, for lung cancer, where tobacco smoking is a major factor for it and other causes of death such as cardiovascular disease.

Breast cancer specific hazards have been estimated in two ways:

1. Regression models. Risk is frequently calculated by combination of a regression model derived from case-control or cohort studies of specific relative risks combined with absolute population-based incidence rates from cancer registries. An early and widely used example of this is the Gail model [32], which has been further developed in the Breast Cancer Surveillance Consortium (BCSC) model [16, 61, 46].

2. Estimation from family pedigree data of the probability of carrying one or more high-risk mutations using segregation analysis, and then using the penetrance of a mutation to alter age-specific risk [18]. Two examples are BRCAPRO and BOADICEA [48, 4].

A novel aspect of the Tyrer–Cuzick model is that it employs both methods [63]. The segregation model estimates the hazard function $h_G(t \mid \boldsymbol{x}_1)$ at age t due to genetic factors, conditional on information (denoted $\boldsymbol{x}_1$) about a woman's family history (family tree) of breast and/or ovarian cancer, and results of any tests in the family from known highly penetrant breast cancer genes *BRCA1* and *BRCA2*. This is then combined with a relative hazard regression function $r(\boldsymbol{x}_2) > 0$ based on other risk factors

$(\boldsymbol{x}_2)$ through

$$(4) \qquad h_1(t \mid \boldsymbol{x}_1, \boldsymbol{x}_2) = h_G(t \mid \boldsymbol{x}_1) r(\boldsymbol{x}_2).$$

We divide family history into two factors $\boldsymbol{x}_1 = \{\boldsymbol{x}_1^{(a)}, \boldsymbol{x}_1^{(b)}\}$ corresponding to (a) breast and (b) ovarian cancer. The vector $\boldsymbol{x}_1^{(\cdot)}$ has components for event time, censoring information and family relationships, and $\boldsymbol{x}_2$ is based on the other risk factors shown in Table 1. Risk models that only account for family history based on a segregation model do not use $\boldsymbol{x}_2$. Regression-only models developed are often proportional-hazards models of the same form as (4), but with age-specific population rates instead of $h_G(\cdot)$; partial information on family history of breast cancer can be included in the regression function, for example, by in-

TABLE 1
Summary of risk factor parameters in model

	Category	Hazard ratio	Mean risk	Reference category
(a) Classic risk factors				
Menopause age (y)	per 5y	1.14	1.08	45–49y
Menarche age (y)	<11	1.16	1	13y
	11	1.07		
	12	1.07		
	13	1		
	14	0.98		
	15	0.93		
	16	0.88		
	17 or older	0.81		
Height (m)	<1.6	1	1.1	<1.6 m
	1.6–1.7	$1.05 + 2 * (\text{height} - 1.6)$		
	1.7 or taller	1.24		
Body mass index	<21	1	1.24	<21
(kg/m^2)	21 to <23	1.14		
(post-menopausal only)	23 to <25	1.15		
	25 to <27	1.26		
	27 or more	1.32		
Age at 1st	Nulliparous	1	1	Nulliparous
childbirth (y)	<17–19	0.74		
	20–24	0.77		
	25–29	0.87		
	30–34	1.01		
	35+	1.11		
Menopausal	Not current	1	1	Not current
hormone therapy	Estrogen-only (current)	1.4		
	Combined (current)	2		
Benign disease	Nonproliferative/none	1	1	None
	Proliferative (usual type)	2		
	Atypical hyperplasia	4		
	Lobular carcinoma in situ	8		
(b) New in v8				
Breast density residual	Visual asssessment scale	1.4 per SD	1	Average density
(age 40y+)	BI-RADS density	1.4 per SD		(age and BMI
	Volumetric percentage	1.4 per SD		adjusted)
SNPs	Continuous	Input	1	Average woman

SD, standard deviation; SNP, single nucleotide polymorphism risk score.

cluding a covariate for the number of affected first-degree relatives.

2.2 Tyrer–Cuzick Segregation Model

2.2.1 *Statistical model.* The genetic risk of breast cancer is modelled through estimates of mutations in *BRCA1* and *BRCA2* genes and an unknown dominant gene. Some of the genes which make up the unknown component have been discovered [28], but most have not, and are only inferred. Even the known ones are rare and individually do not make a major contribution, and are not so commonly tested for even in women with a family history. We use the variable c_1 for *BRCA1/2* status (0 if not a *BRCA1/2* carrier, 1 if a *BRCA1* carrier, and 2 if *BRCA2* carrier; joint carriers are extremely rare and are modelled as *BRCA1* carriers). The unknown dominant gene is denoted c_2 (0, 1 if respectively, not a carrier or a carrier). The model makes the following assumptions.

A.i Breast $x_1^{(a)}$ and ovarian cancer $x_1^{(b)}$ family histories are conditionally independent given c_1 and c_2:

$$p(x_1 \mid c_1, c_2) = p\{x_1^{(a)} \mid c_1, c_2\} p\{x_1^{(b)} \mid c_1, c_2\},$$

where $p(\cdot)$ is general notation to denote a probability density or mass function.

A.ii c_1 and c_2 are independent: $\{p(c_1, c_2) = p(c_1) \times p(c_2)\}$.

A.iii The unknown gene c_2 is not associated with ovarian cancer $[p\{x_1^{(b)} \mid c_1, c_2\} = p\{x_1^{(b)} \mid c_1\}]$.

A.iv The penetrance $S_G(t \mid c_1)$, that is, the probability of developing disease by age t, is known for breast and ovarian cancer in *BRCA1/2* carriers. The prevalence $p(c_1)$ of *BRCA1*, *BRCA2* and the unknown gene are known in the population.

A.v A proportional-hazards model is assumed for the incidence associated with the unknown gene using two parameters $\theta = (\beta, \gamma)$ through

$$(5) \qquad S_G(t \mid c_1, c_2; \theta) = S_0(t \mid c_1)^{\exp(\gamma c_2)},$$

where γ quantifies the log relative hazard attributable to the unknown gene, β is its prevalence (the per-allele proportion is $1 - \sqrt{1 - \beta}$), and $S_0(t \mid c_1)$ is a baseline survivor function. From assumption (A.iv) and equation (5) the baseline survivor function is obtained by solving

$$
(6) \qquad
\begin{aligned}
S_G(t \mid c_1; \theta) = {} & (1 - \beta) S_0(t \mid c_1) \\
& + \beta S_0(t \mid c_1)^{\exp(\gamma c_2)},
\end{aligned}
$$

for example, via a Newton-Raphson iteration [63].

A.vi Mendelian inheritance, that is, there is a 50% chance that one of the two alleles from each parent is inherited by their offspring.

From these assumptions the breast cancer risk conditional on x_1 is obtained as

$$
(7) \qquad
\begin{aligned}
S_G(t \mid x_1; \theta) = {} & \sum_{c_1=0}^{2} \sum_{c_2=0}^{1} S_G(t \mid c_1, c_2; \theta) \\
& \times p(c_1, c_2 \mid x_1; \theta),
\end{aligned}
$$

where the weights $p(c_1, c_2 \mid x_1; \theta)$ in (7) are obtained from the following application of Bayes' rule. Denoting $d = (c_1, c_2)$ then

$$(8) \qquad p(d \mid x_1; \theta) = \frac{p(x_1 \mid d; \theta) p(d; \theta)}{p(x_1; \theta)}.$$

$p(d; \theta)$ uses assumptions given in (A.ii, A.iv, A.v), and $p(x_1 \mid d; \theta)$ is obtained following A.vi (described in more detail in [63]), first separately for breast $x_1^{(a)}$ and ovarian cancer $x_1^{(b)}$ family histories and then they are combined through (A.i). When gene testing has been done for *BRCA1/2* the values for c_1 are taken from this, and the familial contribution only applied to the unknown gene c_2. Most of the computer code in the Tyrer–Cuzick algorithm is involved in calculating the likelihood $p(x_1 \mid d; \theta)$.

2.2.2 *Parameter estimates.* Some changes to the segregation model have been made since the original model [63], starting from version 7 onwards. *BRCA1/2* prevalence and penetrance are now from the 1950+ birth cohort estimates reported in [4]. The penetrance estimate is now lower than used in original model [63] (previously taken from [31]), which was attributed in [4] to bias in the original estimate [31] resulting from a focus on families with multiple cases of breast cancer. Prevalence of *BRCA1* is taken to be 0.06% and for *BRCA2* is 0.10% (previously 0.11% and 0.12%, respectively). Smoothed first breast cancer rates from the Thames Registry 2005–2009 were used to calibrate overall risk, based on data reported by 5y age group and smoothed by loess (with the smoother 'span' parameter 0.2 chosen by eye). The unknown gene calibration has not been altered since the model was first introduced ($\beta = 11.4\%$ and $\exp(\gamma) = 13.04$, fitted to data from [3]; see [63]). The impact of the unknown gene depends mostly on the product $\beta \exp(\gamma)$ and it is difficult to accurately estimate these two parameters separately. The best fit gave a very large value for $\exp(\gamma)$ but almost equally good fits would arise if γ was substantially reduced and β increased accordingly. Attributing risk to a single gene is a simplification, as risk is likely to reflect a combination of genes. However it is clear that the contribution of this unknown gene is greater than that for *BRCA1* or *BRCA2*, any other known gene or combination of single nucleotide polymorphisms, and treating this either as a single unknown gene or a polygenic risk score has minimal impact on its predictive value.

2.3 Regression Model

2.3.1 *Model*. The relative hazard $r(x_2)$ in model (4) is normalised by the mean risk in the population. This requires information on the relative hazard associated with other nonfamilial (personal, hormonal and lifestyle) risk factors {denote this function $\phi(x_2) > 0$} and their population prevalence {$f(x_2)$ where $\int f(x_2)\,dx_2 = 1$} in order to obtain the normalised relative hazard

$$(9) \qquad r(x_2) = \frac{\phi(x_2)}{\int \phi(x_2) f(x_2)\,dx_2}.$$

In the Tyrer–Cuzick model this is approximated by treating each risk factor as independent (breast cancer risk factors in the model are largely independent of each other) and taking

$$(10) \qquad r(x_2) = \prod_{j=1}^{q} \left\{ \frac{\phi(x_{2j})}{\int \phi(x_{2j}) f(x_{2j})\,dx_{2j}} \right\}.$$

When the value of a risk factor is unknown it can be left blank and the normalised relative hazard associated with that factor is taken to be 1 (woman assumed to have population risk). The mean risk constants used in the model are shown in Table 1.

2.3.2 *Parameters*. The relative hazards and prevalence of risk factors in the original model were reported in [63], and included in Table 1. There have been changes and the current values are described below.

Atypical hyperplasia or LCIS The original model treated atypical hyperplasia or lobular carcinoma insitu (LCIS) as an independent risk factor. Evidence in [25] and [9] indicated that other risk factors should not be included when a women has been diagnosed with atypical hyperplasia as it is more of an intermediate endpoint. As a result, we modified the model to take the maximum risk of atypical hyperplasia (only), LCIS (only) and family history risk combined with other risk factors (based on the 10y risk). For example, a young *BRCA1/2* carrier with atypical hyperplasia would not have her risk assessment modified by atypical hyperplasia. This is likely to be conservative, but currently there is inadequate data to model the joint effects of atypia and other risk factors accurately.

Unknown benign disease In general, the risk of subsequent breast cancer associated with a benign lesion depends on the histology, with no increased risk associated with nonproliferative lesions, about a two-fold risk associated with hyperplasia of the usual type and about a 4-fold risk associated with atypical hyperplasia [27]. Based on prevalence data reported by [27], the relative hazard from benign disease when a women has had a biopsy but pathology is unknown is taken to be 1.3; this option was not available originally. Nonbiopsied lesions are not considered in the model.

Menopause hormone therapy The model assumes a maximum relative hazard of 2.0 for combined therapy with an oestrogen and progestin and 1.4 for oestrogen-only hormone therapy. The relative hazard begins at unity in year 1, is half the maximum excess relative hazard in year 2 and then the maximum relative hazard therafter until stopping therapy. Risk is then ramped down following cessation to be 2/3 of the maximum excess in the first year after stopping, then one third in the next year, and no increase thereafter. The relative risk is adjusted if BMI (body mass index) is known, being decreased by 10% of the average excess risk if obese (BMI $\geq$ 30 kg/m^2), and increased by 10% if BMI <25 kg/m^2 (not overweight or obese). These assumptions are based on results from the Women's Health Initiative trial [17]) and the Million Women Study [51].

2.4 New Additions to the Model

2.4.1 *Mammographic density*. Breast density is correlated with some of the other risk factors in the model, most strongly age and body mass index. It is also measured in different ways that require different calibration. We allow three different ways to input density: a visual assessment in which (ideally) two readers estimate the percentage of the breast covered by dense (opaque) tissue and the average is used, the BI-RADS system which corresponds to four categories, and an automated volumetric system (Volpara) that estimates the percentage volume of dense tissue based on the radiographic intensity of each pixel. In order to incorporate breast density into the model without changing the effect sizes of the other factors, we first developed a measure of breast density which is independent from age and BMI at mammogram, by taking the difference between observed and expected density given age and BMI. This was developed for the different measures in case-control studies, where expected density was modelled in controls by fitting a generalized additive model [70] of different breast density measures against splines for age and BMI (see Figure 1 for a fit to volumetric percent density). The risk associated with density was then calibrated by estimating the effect of the density residual, adjusted for other factors in the Tyrer–Cuzick model using case-control and cohort studies [68, 15, 14].

2.4.2 *Polygenic risk scores*. Numerous studies have demonstrated the utility of polygenic risk scores. These use common gene variants which individually carry small risks but as a group provide useful information. Their utility has been observed in average risk women [66, 46] and in high-risk women [22, 30], and they combine effectively independently with classical risk factors and breast density [64, 65]. Currently, the breast cancer association consortium has validated more than 170 risk-modifying SNPs at genome-wide significance [44], and this is likely to increase still further.
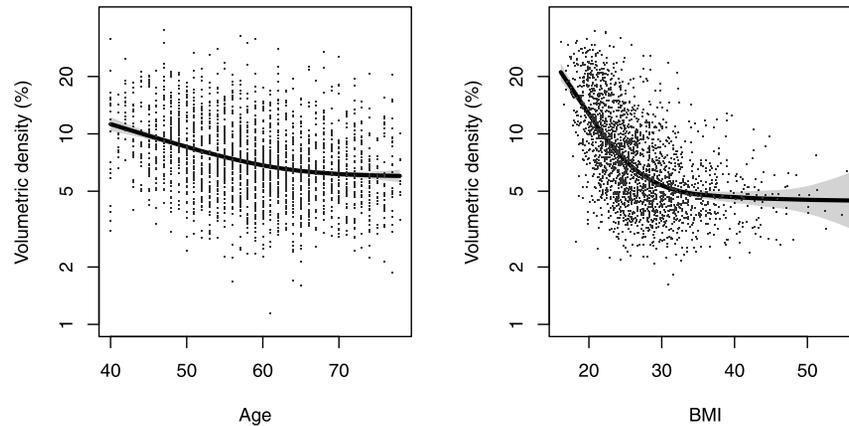
FIG. 1.   *Volumetric percentage density vs age and body mass index (BMI), with joint nonparametric smooths (line —). Taken from* [12].

To provide an overall relative risk estimate from a collection of independent SNPs polygenic risk scores may be obtained by using published data on their per-allele risks, and allele frequencies. First, the odds ratio for each of the three SNP genotypes is determined (no risk alleles, 1 risk allele, and 2 risk alleles) and these are normalised to obtain an average of unity using reported risk allele frequencies. In our model, we assume the odds of two copies of the abnormal gene is the square of that for one copy, and their distribution follows the Hardy–Weinberg law. Finally, to obtain an overall polygenic risk score the odds ratios for each of the genotypes are multiplied together assuming independence [46].

Data show very small correlations between polygenic risk scores and classical risk factors. We have observed that an accurate fit is obtained by treating the score as independent of the other factors in the Tyrer–Cuzick model [22, 30, 65]. However, theoretically, a polygenic risk score will explain some of the genetic aggregation modelled through the segregation model, and so this should be down weighted in the risk assessment. Work to refine this component is ongoing.

3. COMPARISON WITH OTHER RISK MODELS

It is important to understand differences between risk models, both for use of the models and to refine the scientific process of improving risk assessment. Here we note some differences between the Tyrer–Cuzick and other models.

3.1 Segregation Models

Some segregation models, including BRCAPRO [48], calculate breast cancer risk only based on estimating the likelihood of carrying major genes based on a personal and family history of breast, ovarian or other cancers related to these genes. The main aim of such models is to estimate risk of being a *BRCA1/2* mutation carrier. They are

mostly used for women with a strong family history in order to determine eligibility for *BRCA1/2* testing. They incorporate both pedigree information including age at onset and results of genetic testing for *BRCA1* and *BRCA2* in the women and their relatives. Such models are less suitable for risk assessment of breast cancer in the wider population, partly because they do not allow for other unknown genetic factors in the model.

The BOADICEA model accounts for other rare but highly penetrant genes (*PALB2*, *CHEK2*, *ATM*), and uses a polygenic model to allow for (unobserved) risk due to other genetic factors than those measured [42]. Define $c_1^* = 0$ if not positive for one of the genes, $c_1^* = 1$ if test positive for *BRCA1*, $c_1^* = 2$ if test positive for *BRCA2* (but not *BRCA1*), $c_1^* = 3$ if test positive for *PALB2* (but not *BRCA1/2*) and so on for next *CHEK2* and last *ATM*. Thus, very rare individuals who test positive for more than one gene are assigned a group $c_1^* = 0, \dots, 5$ based on a hierarchy. The model has the following form for the breast cancer specific hazard:

$$h_1\{t \mid c_1, u(t)\}$$
$$= h_{1B}(t) \exp\left\{ \sum_{j=1}^{5} \beta_j(t) I(c_1^* = j) + u(t) \right\},$$

where t is age, $h_{1B}(t)$ is a baseline hazard function, $I(\cdot)$ the indicator function, $\beta_j(t)$ are time-dependent relative hazards for each genotype, and $u(t)$ is a time-dependent random effect from a normal distribution with mean zero and standard deviation σ_t. When the woman has not been personally tested, the probability of being a carrier is estimated based on the pedigree and testing results (if any) in family members. The random (polygenic) effects arise from a combined effect of several unmeasured and largely unknown genes, each of which contributes a small effect. Large scale genome-wide association studies have identified that the polygenic model for genetic susceptibility is

partly correct, but a large proportion of familial aggregation of risk is still not explained by known major genes or polygenic risk [44].

For most pedigrees arising in the general population one might expect only small differences for risk estimates from polygenic vs. single unknown gene models. Qualitatively, the unknown single dominant gene model is a good approximation to the polygenic model: the genotype is never observed and so gives rise to a range of effects through equations (7) and (8). Indeed, one study that compared the fit between the unknown dominant gene and polygenic models did not show an appreciable difference between them when considering additional genetic factors other than *BRCA1/2* [5].

Another difference between BOADICEA and Tyrer–Cuzick is that the hazard for the unknown genetic effect γ is assumed to not vary with age in the Tyrer–Cuzick model, whereas the polygenic variance varies with age in the BOADICEA model. Thus, the BOADICEA model leads to a larger decline in the relative risk as the age of the affected relative increases and appears to fit better with the literature. We will update this aspect of the model in the next version.

3.2 Regression Function Models

Several models have been developed based on a regression function only. These include the Gail (or BCRAT) [32], the Breast Cancer Surveillance Consortium (BCSC) [61, 62] and Rosner–Colditz models [52, 72]. Neither the Gail nor BCSC model include all the risk factors included in the Tyrer–Cuzick model, but they do account for ethnic differences. The Rosner–Colditz model includes the terms for the same broad risk factors as in the Tyrer–Cuzick model, and some additional factors including alcohol consumption, adolescent body somatotype (shape/size) and hormone levels. Assumptions on the risks and prevalence of risk factors differ between the models, as do the use of interaction terms.

The Tyrer–Cuzick model includes an interaction between menopausal status and BMI, assuming it is only a postmenopausal risk factor. Some data suggest that BMI may actually be a protective factor for premenopausal breast cancer but this is not included in the model [58]. Interactions are also included for BMI and HRT (as discussed above), and atypical hyperplasia (or LCIS) and all other risk factors in that it is treated as an intermediate endpoint, and the maximum risk for it alone or all other factors combined is used, following analysis from [9]. No other interaction terms are currently included.

The Gail model includes an interaction between (i) ethnicity and all other risk factors, and for certain ethnic groups an interaction between (ii) age and number of biopsies, and (iii) age at first child and number of affected first-degree relatives (negative interaction). For example,

both latter interactions are included for white women, but they are not used for Asian women.

The BCSC model includes interactions between a woman's age with (i) ethnicity, (ii) benign disease, (iii) family history (negative interaction) and (iv) breast density (negative interaction), by allowing for different regression coefficients in 10y intervals starting at age 40y [62].

The Rosner–Colditz model allows for interactions between: (i) BMI and duration of premenopausal period (time from menarche to start of menopause) (ii) BMI and current duration of menopausal interval, (ii) alcohol after menopause and hormone replacement therapy, (iii) benign disease with age at menarche, (iv) benign disease with duration of premenopause, (v) benign disease with duration of menopause, (vi) height with duration of premenopause (vii) height with duration of postmenopause.

All the models (except Tyrer–Cuzick) include interactions relating to benign disease, but their forms are quite different. In the Gail model, for white women the relative risk of previous biopsies is less in women older than 50y than for younger women. In the BCSC cohort used to fit their model [62] this pattern was not observed for proliferative disease, and for nonproliferative disease the direction was even reversed (a larger effect for women older than 50y). In the Rosner–Colditz model, there is a positive interaction with age at menarche and a negative interaction for the other terms [52], suggesting an older age at menarche is not protective for women with benign disease. These three models have been fitted in populations with individual-level data. Relative hazards for white women in the Gail model were estimated using a case-control study with 2842 cases and 3146 controls, including 179 cases with a prior biopsy aged <50y and 487 cases with a prior biopsy aged 50y+ [32]. The BCSC model was fitted to a cohort with more than 1 million women with up to 10y follow up, and included 6204 women with proliferative disease without atypia and 177 subsequent cancers [62]. A recent Rosner–Colditz model update included approximately 100,000 women of whom 5246 had developed breast cancer. Benign breast disease (nonproliferative and proliferative) was reported in approximately 18,000 women at entry [52].

Potential interactions between benign disease and other risk factors have also been investigated in other studies, but also with inconclusive findings. For example, in a cohort from Nashville a negative interaction between nonproliferative disease and age was observed, such that women diagnosed with nonproliferative disease at an older age had lower breast cancer risks than women in the cohort without proliferative disease [27]; this has not been evaluated in any of the four models considered here.

Overall, due to the current lack of strong evidence to support an interaction of benign disease with age, no such

interaction is included in the Tyrer–Cuzick model, and it assumes nonproliferative benign disease does not confer any increased risk.

4. ABSOLUTE CUMULATIVE RISK ASSESSMENT AND CALIBRATION

4.1 Framework

It is important to be aware of the difference between using a model to project risks and calibrating a model from a cohort where follow-up data already exists. In the former case routine data only provides risk assuming no inter-current mortality, and a specific adjustment needs to be made for competing mortality (cf. equation (3)). In the latter case, inter-current deaths are known and can be treated as part of the censoring process. This has led to different methods for computing absolute cumulative risk, which are essential for interpreting the model predictions, and subsequently to assess how well the model is calibrated to the population under study. This has not been widely appreciated, and can lead to different estimated values of expected absolute risk from the same disease model. In this section, we review some different methods for computing expected cumulative risk under the following setup.

Assume there is a sample of $i = 1, \ldots, n$ independent individuals with data on risk factors. The risk model evaluates each risk from current age t_0 to age t, defined as $P_{x_i}(t_{0i}, t)$ or $P_i(t)$ using shorthand notation following equation (3). The time to breast cancer or death is subjected to right-censoring, so that one observes $\check{T} = \min(T_1, T_2, T_C)$, where T_1 is time to disease, T_2 is the time to death from other causes and T_C is a right-censoring time.

An issue to contend with for computing the expected probability of breast cancer is that the 'at-risk' interval should be from t_0 to T_C, whereas $\check{T}$ is often used. T_C is not observed when $\min(T_1, T_2) < T_C$ and this presents difficulties as its (conditional) distribution can be hard to determine. Calculating the expected risk using probability of breast cancer over $(t_0, \check{T})$ is liable to underestimate expected risk. A solution is to use lifetable methods which are based on the sub-hazard of breast cancer and not the probability of breast cancer, but only need to be computed over $(t_0, \check{T})$. We next consider these different methods in more detail.

4.2 Expected Risk Based on the Cumulative Disease-Specific Hazard Function

The expected number of breast cancer cases in the cohort may be obtained by first integrating breast cancer specific hazards over the observed follow-up period, to obtain each individual's cumulative hazard for breast cancer during the period she is at risk. Formally, let $Y_i(t) = I(\check{T}_i > t)$ denote the at-risk process for individual

$i = 1, \ldots, n$, and $h_1(t \mid x_i) = h_{1i}(t)$ the disease-specific hazard function sub-model. Then the expected number of individuals with disease $(J = 1)$ is

$$
(11) \quad
\begin{aligned}
E^{(H)} &= \sum_{i=1}^{n} \int_{t_{0i}}^{\infty} h_{1i}(u) Y_i(u) \, du \\
&= \sum_{i=1}^{n} H_{1i}(\check{T}_i),
\end{aligned}
$$

where

$$
H_{1i}(t) = \int_{t_{0i}}^{t} h_{1i}(u) \, du.
$$

Equation (11) may be adapted to compute the expected number of deaths from other causes by replacing h_1 with h_2, etc. Indeed, by letting h_2 reflect all-cause mortality (including breast cancers) this is equivalent to the time honoured lifetable method to compute the expected number of deaths [40]. Thus, while it might seem surprising to base the expected number of women diagnosed with breast cancer on the disease-specific cumulative hazard rather than the smaller estimate of the probability of breast cancer, this method is firmly established in lifetable and other settings (see, e.g., the log-rank test).

4.3 Expected Risk Based on Cumulative Incidence

The previous section is the preferred method (see below) but another approach is to calculate expected risk using the model probability of disease allowing for death from other causes (cumulative incidence). This is most straightforward when the model is used in a cohort with no censoring from starting age t_0 up to age t. In this case, the expected number is a summation of the conditional cumulative incidence in (3):

$$
(12) \quad E^{(P)}(t) = \sum_{i=1}^{n} P_i(t),
$$

where $P_i(t) \equiv P_x(t_0, t)$ for the ith individual. It is extremely uncommon to have such a dataset, and the calculation is more complicated under the most common scenario of right censoring because we need an estimate of $P\{J = 1, \min(T_1, T_2, T_C) \leq t \mid x_i\}$, whereas $P_i(t)$ estimates $P\{J = 1, \min(T_1, T_2) \leq t \mid x_i\}$ (see equation (3)). We consider three types of censoring processes for prospective risk calculation.

First, if censoring is fixed to be to a common follow-up time for each patient $t_i = t_{0i} + A$, such as $A = 5$ years, then

$$
(13) \quad E^{(P)}(A) = \sum_{i=1}^{n} P_i(t_{i0} + A).
$$

Second, when the censoring time T_C is variable between individuals but it is deterministic, such as being due to

variable calendar time enrolment and a common follow-up for all i then

$$(14) \qquad E^{(P)} = \sum_{i=1}^{n} P_i(t_{i0} + T_{Ci}).$$

A third case is when the potential censoring time is unknown but is a stochastic process with known distribution. Then one possibility is to base expected risk on

$$
\begin{aligned}
(15) \quad &P_C(t) \\
&= \int_{t_0}^{t_0+t} h_1(u \mid \boldsymbol{x}) \\
&\quad \times \exp\Bigg[-\int_{t_0}^{t_0+u} \{h_1(v \mid \boldsymbol{x}) \\
&\qquad\qquad + h_2(v \mid \boldsymbol{x})\}\,\mathrm{d}v\Bigg] S_C(u \mid \boldsymbol{x})\,\mathrm{d}u,
\end{aligned}
$$

where S_C is the survivor function for the censoring distribution. Then

$$(16) \qquad E^{(P)} = \sum_{i=1}^{n} P_{Ci}(t_{i0} + \check{T}_i).$$

If S_C is unknown then one will need to estimate it, for example by assuming censoring is independent of risk factors $\boldsymbol{x}$ and using the Kaplan–Meier estimator [41].

4.4 Calibration

The above methods based on cumulative disease-specific hazards or cumulative incidence may be used to assess calibration by comparison of the observed number of women with disease (O) with expected (E), both overall and in sub-groups [38]. If it is assumed that disease is rare (this assumption is relaxed in Section 4.7), then a test may be constructed assuming that O is generated from a Poisson distribution, whose rate would be E if the model was perfectly calibrated. Exact Poisson confidence intervals on O may be used to compute a confidence interval for O/E (treating E as fixed) and determine if it covers unity [33].

4.5 Some Biased Estimates of Absolute Risk and Their Effect on Calibration

Some commonly-used methods are biased, but this can be overcome using hazard-based methods where the appropriate 'at-risk' interval is used. We next consider three methods that yield biased estimates of E, the expected number of cancers, which in turn leads to biased estimates of the observed to expected ratio O/E.

1. Under right censoring some studies [19, 24, 7, 55] have used cumulative incidence as the measure of expected risk, but estimated it via

$$(17) \qquad \sum_{i=1}^{n} P_i(\check{T}_i).$$

This measure of expected risk is biased towards zero because it only computes risk until the event time $\check{T}$. For example, suppose we wish to compute the expected number of deaths in a sample of babies over the next 200y. If the mortality model is well calibrated but risk of death until the age at which each person actually dies is summated (as equation (17)), then fewer deaths will be expected than the number of babies in the sample. The correct analysis based on cumulative incidence would summate risk of death by age 200y for each individual (cf. equation (12)). In mitigation, the bias will be small, and will have a minimal impact on the conclusions drawn when the event is rare.

2. In some studies it has been reported [2] that the expected number with disease is computed as

$$(18) \qquad \sum_{i=1}^{n} [1 - \exp\{-H_{1i}(\check{T}_i)\}],$$

that is, summating 1 minus the net survival (the estimand for Kaplan–Meier estimation) rather than (11). In general, this is biased towards zero for the expected number of breast cancers because the term in the summation is less than or equal to the cumulative hazard $H_{1i}(\check{T}_i)$. However, this is a small bias if the cumulative hazard H is small since $1 - \exp(-H) = H + O(H^2)$. Also note that $1 - \exp\{-H_{1i}(\check{T}_i)\} \geq P_i(\check{T}_i)$, so is less biased than summating $P_i(\check{T}_i)$ (i.e., equation (17)).

3. Finally, calibration of breast cancer risk has sometimes been assessed using a fixed follow-up time for all by only including those who could be followed up at least that long [61, 54]. This in inefficient because some data would be excluded, and it should be used with caution when there is censoring. For example, suppose we seek to assess calibration of 5y risk when there is right censoring in the data, and include only those without breast cancer who were not censored by 5y and all breast cancer cases diagnosed up to 5y. If the model is correct (and censoring is noninformative), then O/E based on (12) will be biased. E will be too small relative to O because some of the noncases will have been excluded from expected risk due to censoring, but they should have contributed to E as they would have been counted as cases if disease had occurred in them.

4.6 An Unbiased Estimate of Expected Risk

To avoid bias, we recommend using $E^{(H)}$ to obtain the expected number with disease. This is based on the cumulative hazard rather than probability of disease (cumulative incidence) and has several advantages compared to $E^{(P)}$ when censoring has been handled appropriately as described in Section 4.3.

First, comparing $E^{(H)}$ with O is a test of whether the disease-specific hazard model is correct. The same method may also be used to test the competing-mortality

model. Comparing $E^{(P)}$ with O is a test of whether the model for cumulative incidence based on combining the disease risk and competing mortality is correct, that is, whether the marginal risk based on (3) is well calibrated. This might require accurate knowledge and data on the competing risk process, which is often not available. It is uncommon for the competing mortality model to be as well developed as the disease risk model, which is of primary interest to the epidemiologist. If the competing mortality model is incorrect, but the disease model is correct, then the expected risk will be incorrect due to an inadequate model for competing mortality. Conversely, even if the cumulative incidence function risk is well calibrated, it is possible that the competing mortality and disease model are both wrong in different ways, and just happen to cancel each other out. Comparing the hazard submodels separately would identify such issues.

Second, when there is stochastic censoring equation (11) is easier to apply than (16) because it does not require an assumption or estimate for the conditional censoring distribution.

Third, it can be a more powerful test. For example, in an extreme example of 200y follow-up on mortality in babies, then one would not be able to show that a model that assigns a constant and excessively high mortality rate was incorrect; whereas analysis based on summating the cumulative hazard of that model to the time of each death would yield an expected number of deaths in excess of observed.

4.7 Extensions

We end this section by describing methods to assess calibration not only for the total population, but also in subgroups. For example, when the data are split into ten groups by decile of the predicted risk at baseline then a test of calibration (more generally model fit) across the groups would be akin to the Hosmer-Lemeshow test for binary outcomes [36, 38]. To extend this strategy to the cumulative incidence framework, we construct $k = 1, \ldots, K$ distinct groups, and assess whether the observed risk matches expected in each of the groups. Then, the test statistic

$$\chi_{K-1}^2 = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k},$$

where the expected E_k are obtained as above, will be approximately χ^2 with $K-1$ degrees of freedom under the null hypothesis of correct prediction and calibration.

This hypothesis testing approach has limitations. First, power to reject inaccurate calibration is limited by the large number of degrees of freedom. Second, with a large enough sample size we would expect to reject model calibration because all models are wrong, and the approach does not directly indicate where the inaccuracies lie.

An alternative approach is to view testing calibration as special case of Poisson regression. Let O_i denote whether individual $i = 1, \ldots, n$ in the sample has disease, and E_i be the expected risk based on the model. Then consider a Poisson regression for

$$\mathrm{E}(O_i \mid E_i) = \exp\{\gamma_0 + \gamma_1 \log(E_i)\}$$
$$= \theta E_i^{\gamma_1},$$

where $\theta = \exp(\gamma_0)$ and γ_1 are unknown parameters. Setting $\gamma_1 = 1$ then the maximum likelihood estimate $\hat{\theta} = \sum_i O_i / \sum_i E$ provides an overall 'calibration-in-the-large' parameter for the entire cohort [21]. When γ_1 is also estimated then it can be used to test for calibration across the range of expected risk (on 1 df), as well as yielding an estimate of how closely the regression line matches observed ($\gamma_1 = 1$ being ideal; 0 being not at all).

Finally, we consider when the disease is relatively common, perhaps because focus is on a high risk group with long follow up. In this case, the Poisson distribution no longer approximates a binomial distribution. However, Poisson regression may be valid if time is broken into shorter segments where the rare disease assumption holds, for example by year, with each segment treated as a 'new' observation in the Poisson regression analysis. This approach is also useful when assessing time-dependent calibration, and which we proceed to consider next.

5. TIME-DEPENDENT CALIBRATION

The focus of many studies has been on whether the cumulative number with disease over a single follow-up period matches the expected number. While important, it does not assess differences through time, such as nonproportionality of the observed to expected risk. In this section, we consider techniques to assess calibration across follow-up time, and develop new graphical methods.

5.1 Methods

One approach is simply to look at the observed and expected number of events $N_j(t)$ for each cause j as a counting process. The observed number is

$$(19) \qquad \hat{N}_j(t) = \sum_{\check{T}_i \leq t} I(J_i = j).$$

The expected number, based on the cumulative hazard approach (as Section 4.2), is

$$
(20) \qquad
\begin{aligned}
N_j(t) &= \sum_{i=1}^{n} \int_0^t Y_i(u) h_{ji}(u) \, \mathrm{d}u \\
&= \sum_{\check{T}_i \leq t} H_{ji}(\check{T}_i) + \sum_{\check{T}_i > t} H_{ji}(t),
\end{aligned}
$$

so that we have extended (11) to be a function of t. One may compare observed $\hat{N}_j(t)$ vs. expected from (20) through follow-up time t, such as is sometimes done to check proportional hazards [6, 34]. Cumulative pointwise (such as from entry to time t), or interval (such as number in a year) confidence intervals may be constructed using exact Poisson confidence intervals, or used to undertake hypothesis tests of calibration up to or near time t.

An issue with the comparisons based on $N(t)$ is that they depend on the sample size and censoring distribution, so cannot be readily used to compare between studies. To circumvent this, one might consider the comparing the mean cumulative hazard among those at risk, that is, for $t < \max_i(\breve{T}_i)$ comparing

$$(21) \qquad \hat{H}_j(t) = \sum_{\breve{T}_i \leq t} \frac{I(J_i = j)}{Y_+(\breve{T}_i)}$$

with its expected value

$$(22) \qquad H_j(t) = \sum_{i=1}^{n} \int_0^t \frac{Y_i(u) h_{ji}(u)}{Y_+(u)} \, du,$$

where the total number at risk at time t is $Y_+(t) = \sum_{i=1}^{n} Y_i(t)$. Note that equation (21) is the Nelson–Aalen estimator of the cumulative hazard [41]. Equation (22) is the expected cumulative hazard based on the breast cancer risk model and observed at-risk process. Time-dependent assessment of the observed to expected risk can be based on $\hat{H}_j(t)$ vs. $H_j(t)$ across follow-up time t. Here we treat $H_j(t)$ as fixed and use the usual confidence intervals for the Nelson–Aalan estimator $\hat{H}_j(t)$ at different times, from which the hypothesis of calibration at a particular time t can also be assessed.

Other functionals of $\hat{H}_j(t)$ and $H_j(t)$ might also be used for assessment of calibration across time. Arguably the most common functional used in applications is the 'net' risk

$$S(t) = 1 - \exp\{-H(t)\},$$

which is often estimated as 1 minus the Kaplan–Meier estimator [41]. There are two options for calculating the expected risk. The first is a mean from the entire cohort at baseline

$$(23) \qquad S_j^{(A)}(t) = n^{-1} \sum_{i=1}^{n} [1 - \exp\{-H_{ji}(t)\}].$$

The second is to use the mean hazard in those still at risk at time t through

$$(24) \qquad S_j^{(B)}(t) = 1 - \exp\{-H_j(t)\}$$

and (22).

A disadvantage of the cumulative hazard (22) and net risk (23), (24) is that interpretation of the component terms is not straightforward in relation to the observable

risk in equation (2). However, if we weight (21) by the proportion at risk using the Kaplan–Meier estimator $\hat{S}(t)$ for all-cause survival (i.e., the survival endpoint is disease or death), then we recover an estimator of the cumulative incidence function through

$$(25) \qquad \hat{F}_j(t) = \sum_{\breve{T}_i \leq t} \frac{I(J_i = j)}{Y_+(\breve{T}_i)} \hat{S}(\breve{T}_i),$$

and inference may be based on estimates of its variance [41]. Expected risk may be obtained from

$$(26) \qquad F_j^{(A)}(t) = n^{-1} \sum_{i=1}^{n} P_i(t),$$

or the mean expected risk at each t over the n individuals.

5.2 Summary

Our recommendation is to focus on comparisons only involving the disease-specific hazard to assess O vs E through follow-up time t. For exploratory plots, we prefer use of the cumulative hazard comparison of (22) with (21). There are several reasons for these recommendations.

First, as for overall calibration, the method may be used to assess the disease risk and competing mortality submodels separately. The cumulative hazard analysis extends quite naturally a total number of events comparison, but avoids dependence on the censoring distribution.

Second, focusing on the hazard leads to using the observed at-risk process to calculate expected risk, which appears to be a more robust method to assess model calibration than only using a risk assessment estimate on everyone at baseline. For instance, if those censored or lost to follow-up are more likely to be at higher predicted risk at entry, then estimates of their expected risk will differ, and so comparison of expected risk between methods based on using the at-risk-process or everyone at baseline is a way to assess whether the censoring process was associated with predicted risk. In general terms this issue has parallels with relative survival, where one might consider the Ederer-I approach (mean risk) or the Ederer-II approach (use the at-risk process) [56].

Third, a statistical advantage of the cumulative hazard compared to the cumulative incidence function is that it may have a smaller variance because it does not require an assumption or estimation of all-cause survival (equation (25)).

Finally, in the exploratory plots when the aim is to assess calibration of a risk model, then differences between the observed to expected cumulative hazard seem easier to interpret than the other methods. In contrast, the cumulative incidence function weighs the cumulative hazard by all-cause survival. The component terms have a natural interpretation, but the observed to expected comparison

is perhaps less easy to interpret for assessing calibration than cumulative cause-specific hazards, because changes could be due to either lack of calibration in the disease or competing mortality models. Kaplan–Meier estimation transforms the cumulative hazard through $1 - \exp(-H)$, so that differences between large cumulative hazards are downweighted. This makes changes in the calibration plot with time less easy to relate to calibration at the hazard level. For example, it is easier to assess whether the hazard function is constant through time by plotting the Nelson–Aalan estimate of the cumulative hazard, than by plotting 1 minus the Kaplan–Meier estimator.

6. EXAMPLE

6.1 Data

We have previously reported an evaluation using the Tyrer–Cuzick model in a cohort from the State of Washington, USA [14]. Between 1996 and 2013, 132,139 women aged 40–73y completed a risk questionnaire and had a measure of breast density taken at entry. They were followed up beginning 6-months after the entry mammogram to the earliest of diagnosis of invasive breast cancer or censoring. Women were censored due to disenrollment (n = 62,331, 48.2%), end of follow-up (n = 48,317, 37.3%), age 75y (n = 15,827, 12.2%), death from other causes (n = 2328, 1.8%) or ductal carcinoma in situ (n = 637, 0.5%). Only aggregate data on competing risk causes were made available for analysis, so we were unable to apply methods that require individual-level data about the cause of censoring.

6.2 Calibration

A total of 2699 breast cancers were observed (O), and based on our preferred method (equation 11) 2757 were expected (E), yielding O/E 0.98 (95%CI 0.94–1.02). We next consider application of incorrect methods discussed in Section 4.5.

A first biased assessment is cumulative incidence over a 5y fixed-time horizon by using all the cases that were diagnosed in the 5y period and all noncases at risk for 5y or more. It we calculate the expected 5y risk in those who are at risk at 5y then $E = 877$ against an observed $O = 1157$ ($O/E = 1.32$). A second biased method is to summate expected cumulative incidence to the last follow-up for each individual (equation 17), yielding $E = 2605$ and $O/E = 1.04$ (95%CI 1.00–1.08). A third biased method is to follow equation (18) and summate the absolute net risks, which gives expected number $E = 2679$ ($O/E = 1.00$).

It also instructive to repeat the above methods in a high-risk group (> 8% ten-year net risk) where greater differences between the methods to calculate E will occur. The Tyrer–Cuzick model with mammographic density identified 4645 women to be in this group, and 273 breast cancers were subsequently diagnosed. Based on the preferred method (11) we find $E = 349$, so that $O/E = 0.78$

(95%CI 0.69–0.88), indicating that the model over predicted the high risk group. If instead (18) was used then $E = 324$ and $O/E = 0.84$ (95%CI 0.75–0.95). Based on (17) we find $E = 310$ so that $O/E = 0.88$ (95%CI 0.78–1.00), where the 95%CI covers unity. Thus, the correct analysis showed a lack of calibration that would not be seen so clearly with the biased methods.

This analysis shows the method to calculate expected risk can cause practically important differences in interpretation of risk model calibration, particularly for high-risk groups.

6.3 Time-Dependent Calibration Plots

Time-dependent calibration was assessed using the methods in Section 5. We present results from overall model calibration (Figure 2) and in the highest and lowest risk deciles (Supplementary Figures S1 and S2 [13]) following [14].

Overall the model appeared well calibrated throughout follow-up time. As follow-up started at six months, the expected hazard in the first year was much lower than the second year of followup. Figure 2(a) shows the model did not adequately take into account the effect of removal of a pool of cancers diagnosed at entry. However, subsequently the model tracked the observed number quite closely. Figures 2(c), (d) show respectively, the Nelson–Aalan and Kaplan–Meier curves. The corresponding observed to expected plots in Figures 2(e), (f) are virtually identical using the expected risk based on those still at risk. There is also very little difference between the two methods to obtain expected risk using the Kaplan–Meier approach, which indirectly confirms that censoring was unrelated to risk assessment.

The same plots were used to assess high and low risk groups. Supplementary Figure S1 considers women who were in the highest predicted risk decile at entry. As noted in Section 6.2, there is some evidence of over-estimation from the model for this group. However, the observed to expected plots in Supplementary Figure S1(e), (f) show a fairly constant pattern across time, with no discernible difference between Nelson–Aalan or Kaplan–Meier comparisons. Supplementary Figure S2 shows calibration of the bottom decile, where calibration appears reasonable through follow up time, albeit with a suggestion that the model had a tendency to under estimate risks.

6.4 Regression Analysis

The above analysis suggests that the risk model was calibrated overall in the cohort, but tended to overestimate risk in those predicted at highest risk; it over-estimated risk in the first year due to omitting screen-detected cases, but the risk predictions thereafter were stable with greater follow-up time. All these aspects may be jointly tested through a Poisson regression model. The model was fitted with an exponential link function. The log hazard per
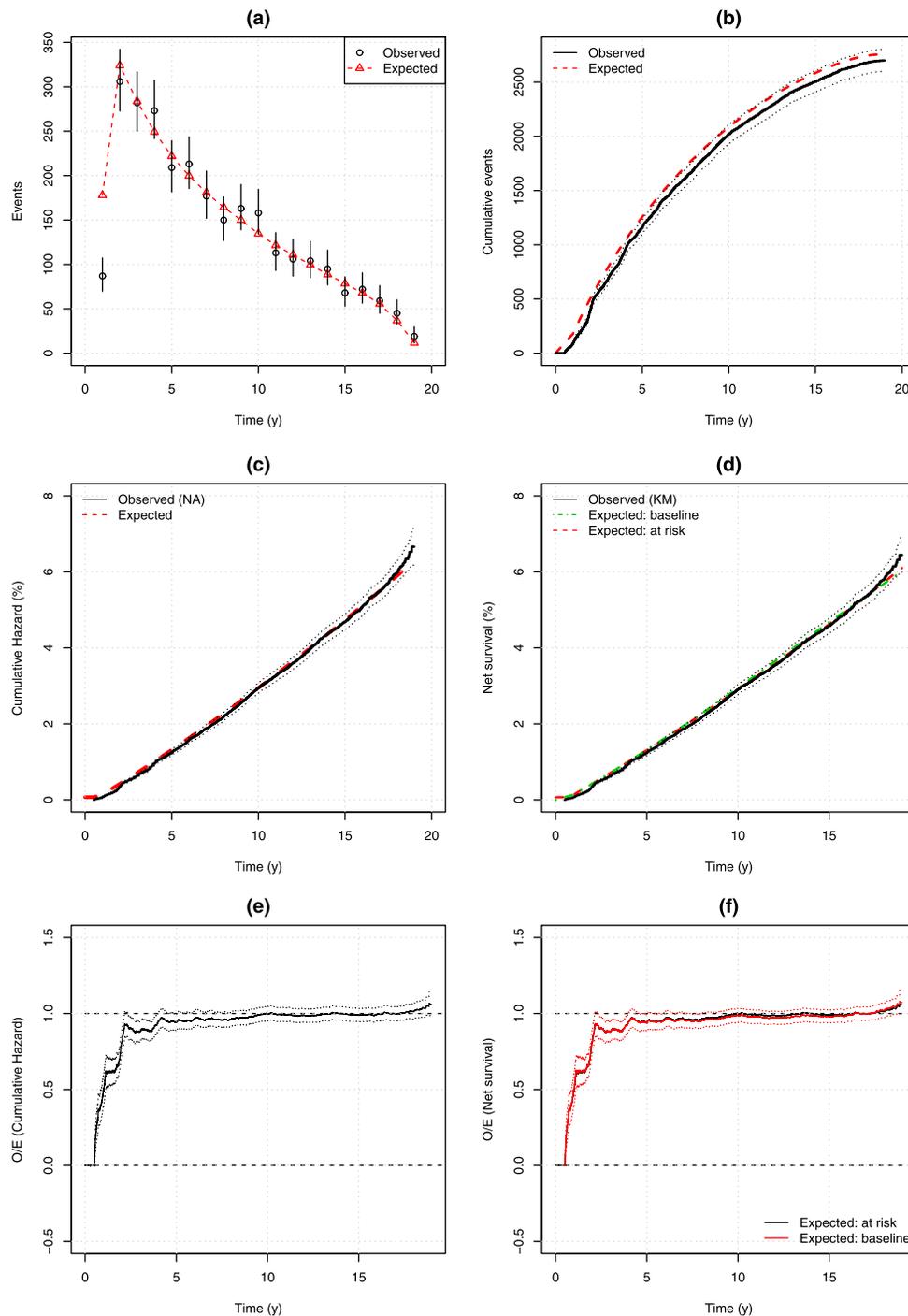
FIG. 2. *Overall calibration of the breast cancer risk model.* (*a*) *Observed* (95%CI) *vs expected number of breast cancers diagnosed for each year of follow-up*, (*b*) *cumulative observed* (95%CI) *vs expected number of breast cancers diagnosed*; (*c*) *observed* (*Nelson–Aalan*, 95%CI) *vs expected cumulative hazards*; (*d*) *observed* (*Kaplan–Meier*, 95%CI) *vs expected* (*obtained via two methods*) *net risks*; (*e*) *Observed divided by Expected cumulative hazard* (95%CI); (*f*) *Observed divided by expected net risk* (*obtained via two methods*) *with 95%CI only for the expected risk based on baseline risk assessment.*

year (or until the event or censoring if in that year) was included as an offset term, and calibration coefficients were estimated for (i) overall lack of calibration in the first year, (ii) stability of calibration from year two onward (follow-up year), and (iii) 10y risk categories from the model as previously defined (<2%, 2 to <3%, 3 to < 5%, 5 to < 8%, ≥ 8%; with 2–3% group as the reference category) [14].

Results are shown in Table 2, taking the exponent of the estimated model parameter to give a calibration coefficient where unity indicates ideal calibration. They confirm the earlier analysis. The risk model over esti-

TABLE 2
Calibration assessment: univariate and Poisson regression (adjusted for variables listed) calibration coefficient estimates with 95% Wald confidence intervals

Term	n	O	E	O/E [univariate]	O/E (95%CI) [adjusted]
Overall (intercept)	132,139	2699	2757	0.98	1.05 (0.94–1.16)
Follow-up time					
Year 1	132,139	87	178	0.49	0.50 (0.40–0.63)
Year 2+ (time)	123,830	2612	2579	1.01	1.00 (0.99–1.01)
10y Risk group					
<2%	53,436	641	548	1.12	1.17 (1.00–1.25)
2 to <3%	33,269	627	603	1.04	1
3 to <5%	29,477	779	784	0.99	0.96 (0.86–1.07)
5 to <8%	11,312	379	473	0.80	0.78 (0.68–0.88)
8%+	4645	273	349	0.78	0.76 (0.66–0.88)

mated risk in the first year by a factor of two (calibration coefficient 0.50, 95%CI 0.40–0.63), which is due to the expected short-term reduction in risk following a negative screen. There was no loss in calibration subsequently (calibration 1.00, 95%CI 0.99–1.01). The highest risk category (10y risk $\geq$ 8%) exhibited evidence of over-estimation of risk relative to the reference average-risk category (calibration overall 0.80, 95%CI 0.69–0.92, i.e., including the intercept), and likewise there was some under-estimation in the lowest-risk group (overall test for calibration across the five groups likelihood-ratio $\chi^2 = 48.6$, df = 4).

7. CONCLUSION

In this paper, we reviewed the statistical foundations of a model to evaluate breast cancer risk at different follow-up times. This model, and corresponding computer program, synthesise findings from many scientific studies by combining risk factors into a hybrid genetic segregation and regression model. The model has been used to guide entry criteria into prevention trials [23], and is recommended by several organisations to guide preventive and early detection strategies, including the American Cancer Society to help determine eligibility for supplemental breast magnetic resonance imaging [57].

We developed new graphical methods for looking at calibration over follow-up time that can be combined with Poisson regression analysis. The impact of using the different methods to assess the calibration of cumulative absolute risk was reviewed. Different methods to account for competing risks when assessing risk model calibration have been used. We note that, given the diversity of methods in use and potential impact on findings, it is important when presenting calibration findings that a detailed description of the method for expected risk is also given [1].

There are opportunities to refine and improve the current risk model. Other factors are known to affect the risk of breast cancer, but most have some difficulties for inclusion in a relatively simple model. Using weight change from age 20 or weight at a young age in addition to current weight does appear to improve risk assessment to some extent [37], but requires accurate recall of previous weight. Alcohol consumption is a well documented risk factor [35], but underestimation of consumption is well known and it is not clear how best to allow for this. The differing roles of weight in the pre- and postmenopausal periods also needs further investigation [58]. Physical activity and dietary factors are also known to affect risk [50, 71], but identification of the key factors and simple accurate measures of these currently present difficulties. Levels of oestrogens and testosterone may also be important especially for postmenopausal women, but these require a blood sample and are not routinely available [72]. Ethnic differences in breast cancer risk are also well documented—notably in the US lower risks in Hispanics and Asians [8]. However, it is unclear as to how these are related to differences in the known risk factors in the model, so would not need model recalibration, or whether there are intrinsic ethnic differences which require use of a separate baseline hazard function, and possibly separate risks for the known factors. Differences between recent migrants and established individuals are also known to exist and complicate any model adjustments [39]. Adding breast density has substantially improved the model, and might be extended by considering longitudinal measurements of density and features in the mammogram other than simply breast density, where more complex algorithms may be able to extract more risk information [67]. The polygenic (SNP) risk score is steadily being improved as more SNPs are added, but this is easily accommodated as all that is required is the relative risk. A bigger challenge is to accurately account for interactions between risk factors. Our model was developed by synthesizing risks for individual factors from large overviews of many studies, and large datasets which contain all the factors

will be needed to fully explore interactions. These are not currently available but are likely to be so soon, as the risk factors used in our model are currently being collected in several large cohorts of patients.

In conclusion, we are gradually moving towards an era of precision medicine, where disease treatment, early detection and prevention strategies will depend on risk assessment. The accuracy of risk models will underpin this approach.

ACKNOWLEDGEMENTS

SUPPLEMENTARY MATERIAL

Supplement to "Risk Models for Breast Cancer and Their Validation" (DOI: 10.1214/19-STS729SUPP; .pdf). Supplementary information.

REFERENCES

[1] ALTMAN, D. G., MCSHANE, L. M., SAUERBREI, W. and TAUBE, S. E. (2012). Reporting recommendations for tumor marker prognostic studies (REMARK): Explanation and elaboration. *PLoS Med.* **9** e1001216+. https://doi.org/10.1371/journal. pmed.1001216.

[2] AMIR, E., EVANS, D. G., SHENTON, A., LALLOO, F., MORAN, A., BOGGIS, C., WILSON, M. and HOWELL, A. (2003). Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme. *J. Med. Genet.* **40** 807–814. https://doi.org/10.1136/jmg.40.11.807.

[3] ANDERSON, H., BLADSTROM, A., OLSSON, H. and MOLLER, T. R. (2000). Familial breast and ovarian cancer: A Swedish population-based register study. *Am. J. Epidemiol.* **152** 1154–1163. https://doi.org/10.1093/aje/152.12.1154.

[4] ANTONIOU, A. C., CUNNINGHAM, A. P., PETO, J., EVANS, D. G., LALLOO, F., NAROD, S. A., RISCH, H. A., EYFJORD, J. E., HOPPER, J. L. et al. (2008). The BOADICEA model of genetic susceptibility to breast and ovarian cancers: Updates and extensions. *Br. J. Cancer* **98** 1457–1466. https://doi.org/10.1038/sj.bjc.6604305.

[5] ANTONIOU, A. C., PHAROAH, P. D. P., MCMULLAN, G., DAY, N. E., STRATTON, M. R., PETO, J., PONDER, B. J. and EASTON, D. F. (2002). A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. *Br. J. Cancer* **86** 76–83. https://doi.org/10.1038/sj.bjc.6600008

[6] ARJAS, E. (1988). A graphical method for assessing goodness of fit in Cox's proportional hazards model. *J. Amer. Statist. Assoc.* **83** 204–212.

[7] BANEGAS, M., GAIL, M., LACROIX, A., THOMPSON, B., MARTINEZ, M., WACTAWSKI-WENDE, J., JOHN, E., HUBBELL, YASMEEN, S. et al. (2012). Evaluating breast cancer risk projections for hispanic women. *Breast Cancer Res. Treat.* **132** 347–353. https://doi.org/10.1007/s10549-011-1900-9.

[8] BANEGAS, M. P., JOHN, E. M., SLATTERY, M. L., GOMEZ, S. L., YU, M., LACROIX, A. Z., PEE, D., CHLEBOWSKI, R. T., HINES, L. M. et al. (2017). Projecting individualized absolute invasive breast cancer risk in US hispanic women. *J. Natl. Cancer Inst.* **109** djw215. https://doi.org/10.1093/jnci/djw215. https://doi.org/10.1093/jnci/djw215

[9] BOUGHEY, J. C., HARTMANN, L. C., ANDERSON, S. S., DEGNIM, A. C., VIERKANT, R. A., REYNOLDS, C. A., FROST, M. H. and PANKRATZ, V. S. (2010). Evaluation of the Tyrer–Cuzick (international breast cancer intervention study) model for breast cancer risk prediction in women with atypical hyperplasia. *J. Clin. Oncol.* **28** 3591–3596. https://doi.org/10.1200/JCO.2010.28.0784.

[10] BOYD, N. F., GUO, H., MARTIN, L. J., SUN, L., STONE, J., FISHELL, E., JONG, R. A., HISLOP, G., CHIARELLI, A. et al. (2007). Mammographic density and the risk and detection of breast cancer. *N. Engl. J. Med.* **356** 227–236. https://doi.org/10.1056/nejmoa062790.

[11] BOYD, N. F., O'SULLIVAN, B., CAMPBELL, J. E., FISHELL, E., SIMOR, I., COOKE, G. and GERMANSON, T. (1982). Mammographic signs as risk factors for breast cancer. *Br. J. Cancer* **45** 185–193.

[12] BRENTNALL, A. R., COHN, W. F., KNAUS, W. A., YAFFE, M. J., CUZICK, J. and HARVEY, J. A. (2019). A casecontrol study to add volumetric or clinical mammographic density into the Tyrer–Cuzick breast cancer risk model. *J. Breast Imaging* **1** 99–106. https://doi.org/10.1093/jbi/wbz006.

[13] BRENTNALL, A. R. and CUZICK, J. (2020). Supplement to "Risk models for breast cancer and their validation." https://doi.org/10.1214/19-STS729SUPP.

[14] BRENTNALL, A. R., CUZICK, J., BUIST, D. S. M. and BOWLES, E. J. A. (2018). Long-term Accuracy of Breast Cancer Risk Assessment Combining Classic Risk Factors and Breast Density. *JAMA Oncology* **4** e180174. https://doi.org/10.1001/jamaoncol.2018.0174.

[15] BRENTNALL, A. R., HARKNESS, E. F., ASTLEY, S. M., DONNELLY, L. S., STAVRINOS, P., SAMPSON, S., FOX, L., SERGEANT, J. C., HARVIE, M. N. et al. (2015). Mammographic density adds accuracy to both the Tyrer–Cuzick and Gail breast cancer risk models in a prospective UK screening cohort. *Breast Cancer Res.* **17** 147+. https://doi.org/10.1186/s13058-015-0653-5.

[16] CHEN, J., PEE, D., AYYAGARI, R., GRAUBARD, B., SCHAIRER, C., BYRNE, C., BENICHOU, J. and GAIL, M. H. (2006). Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. *J. Natl. Cancer Inst.* **98** 1215–1226.

[17] CHLEBOWSKI, R. T. and ANDERSON, G. L. (2011). The influence of time from menopause and mammography on hormone therapy-related breast cancer risk assessment. *J. Natl. Cancer Inst.* **103** 284–285. https://doi.org/10.1093/jnci/djq561.

[18] CLAUS, E. B., RISCH, N., THOMPSON, W. D., CLAUS, E. B., RISCH, N. and THOMPSON, W. D. (1993). The calculation of breast cancer risk for women with a first degree family his-

tory of ovarian cancer. *Breast Cancer Res. Treat.* **28** 115–120. https://doi.org/10.1007/bf00666424.

[19] COSTANTINO, J. P., GAIL, M. H., PEE, D., ANDERSON, S., REDMOND, C. K., BENICHOU, J. and WIEAND, H. S. (1999). Validation studies for models projecting the risk of invasive and total breast cancer incidence. *J. Natl. Cancer Inst.* **91** 1541–1548.

[20] CROWDER, M. J. (2001). *Classical Competing Risks*, 1st ed. CRC Press/CRC, Boca Raton, FL.

[21] CROWSON, C. S., ATKINSON, E. J. and THERNEAU, T. M. (2016). Assessing calibration of prognostic risk scores. *Stat. Methods Med. Res.* **25** 1692–1706. MR3541121 https://doi.org/10.1177/0962280213497434

[22] CUZICK, J., BRENTNALL, A. R., SEGAL, C., BYERS, H., REUTER, C., DETRE, S., LOPEZ-KNOWLES, E., SESTAK, I., HOWELL, A. et al. (2017). Impact of a panel of 88 single nucleotide polymorphisms on the risk of breast cancer in high-risk women: Results from two randomized tamoxifen prevention trials. *J. Clin. Oncol.* **35** 743–750. https://doi.org/10.1200/JCO.2016.69.8944.

[23] CUZICK, J., SESTAK, I., CAWTHORN, S., HAMED, H., HOLLI, K., HOWELL, A. and FORBES, J. F. (2015). Tamoxifen for prevention of breast cancer: Extended long-term follow-up of the IBIS-I breast cancer prevention trial. *Lancet Oncol.* **16** 67–75. https://doi.org/10.1016/s1470-2045(14)71171-4.

[24] DECARLI, A., CALZA, S., MASALA, G., SPECCHIA, C., PALLI, D. and GAIL, M. H. (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** 1686–1693. https://doi.org/10.1093/jnci/djj463.

[25] DEGNIM, A. C., VISSCHER, D. W., BERMAN, H. K., FROST, M. H., SELLERS, T. A., VIERKANT, R. A., MALONEY, S. D., PANKRATZ, V. S., DE GROEN, P. C. et al. (2007). Stratification of breast cancer risk in women with atypia: A mayo cohort study. *J. Clin. Oncol.* **25** 2671–2677. https://doi.org/10.1200/JCO.2006.09.0217.

[26] DE STAVOLA, B. (1987). Statistical facts about cancers on which doctor Rigoni–Stern based his contribution to the surgeons' subgroup of the IV Congress of the Italian scientists on 23 September 1842. *Stat. Med.* **6** 881–884. https://doi.org/10.1002/sim.4780060803.

[27] DUPONT, W. D., DEGNIM, A. C., SANDERS, M. E., SIMPSON, J. F. and HARTMANN, L. C. (2018). Risk Factors for Breast Carcinoma in Women With Proliferative Breast Disease. In *The Breast* 264–271.e2. Elsevier, Amsterdam. https://doi.org/10.1016/b978-0-323-35955-9.00020-9.

[28] EASTON, D. F., PHAROAH, P. D., ANTONIOU, A. C., TISCHKOWITZ, M., TAVTIGIAN, S. V., NATHANSON, K. L., DEVILEE, P., MEINDL, A., COUCH, F. J. et al. (2015). Gene-panel sequencing and the prediction of breast-cancer risk. *N. Engl. J. Med.* **372** 2243–2257.

[29] EASTON, D. F., PHAROAH, P. D. P., ANTONIOU, A. C., TISCHKOWITZ, M., TAVTIGIAN, S. V., NATHANSON, K. L., DEVILEE, P., MEINDL, A., COUCH, F. J. et al. (2015). Gene-panel sequencing and the prediction of breast-cancer risk. *N. Engl. J. Med.* **372** 2243–2257. https://doi.org/10.1056/NEJMsr1501341.

[30] EVANS, D. G., BRENTNALL, A., BYERS, H., HARKNESS, E., STAVRINOS, P., HOWELL, A., NEWMAN, W. G. and CUZICK, J. (2017). The impact of a panel of 18 SNPs on breast cancer risk in women attending a UK familial screening clinic: A case-control study. *J. Med. Genet.* **54** 111–113. https://doi.org/10.1136/jmedgenet-2016-104125.

[31] FORD, D., EASTON, D. F., STRATTON, M., NAROD, S., GOLDGAR, D., DEVILEE, P., BISHOP, D. T., WEBER, B., LENOIR, G. et al. (1998). Genetic heterogeneity and penetrance

analysis of the BRCA1 and BRCA2 genes in breast cancer families. *Am. J. Hum. Genet.* **62** 676–689.

[32] GAIL, M. H., BRINTON, L. A., BYAR, D. P., CORLE, D. K., GREEN, S. B., SCHAIRER, C. and MULVIHILL, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J. Natl. Cancer Inst.* **81** 1879–1886. https://doi.org/10.1093/jnci/81.24.1879.

[33] GAIL, M. H. and PFEIFFER, R. M. (2005). On criteria for evaluating models of absolute risk. *Biostatistics* **6** 227–239. https://doi.org/10.1093/biostatistics/kxi005.

[34] GRØNNESBY, J. K. and BORGAN, Ø. (1996). A method for checking regression models in survival analysis based on the risk score. *Lifetime Data Anal.* **2** 315–328. https://doi.org/10.1007/BF00127305.

[35] COLLABORATIVE GROUP ON HORMONAL FACTORS IN BREAST CANCER (2002). Alcohol, tobacco and breast cancer—collaborative reanalysis of individual data from 53 epidemiological studies, including 58 515 women with breast cancer and 95 067 women without the disease. *Br. J. Cancer* **87** 1234–1245. https://doi.org/10.1038/sj.bjc.6600596.

[36] HARRELL, F. E., LEE, K. L. and MARK, D. B. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15** 361–387. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4%3C361::AID-SIM168%3E3.0.CO;2-4.

[37] HIDAYAT, K., YANG, C. M. and SHI, B. M. (2018). Body fatness at a young age, body fatness gain and risk of breast cancer: Systematic review and meta-analysis of cohort studies. *Obes. Rev.* **19** 254–268. https://doi.org/10.1111/obr.12627.

[38] HOSMER, D. W., LEMESHOW, S. and STURDIVANT, R. X. (2013). *Applied Logistic Regression*. Wiley, Hoboken, NJ, USA. https://doi.org/10.1002/9781118548387.

[39] JOHN, E. M. (2005). Migration history, acculturation, and breast cancer risk in hispanic women. *Cancer Epidemiol. Biomark. Prev.* **14** 2905–2913. https://doi.org/10.1158/1055-9965.EPI-05-0483.

[40] KEIDING, N. (1987). The method of expected number of deaths, 1786–1886–1986. *Int. Stat. Rev.* **55** 1–20. MR0962938 https://doi.org/10.2307/1403267

[41] LAWLESS, J. F. (2003). *Statistical Models and Methods for Lifetime Data*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley-Interscience, Hoboken, NJ. MR1940115

[42] LEE, A. J., CUNNINGHAM, A. P., TISCHKOWITZ, M., SIMARD, J., PHAROAH, P. D., EASTON, D. F. and ANTONIOU, A. C. (2016). Incorporating truncating variants in PALB2, CHEK2, and ATM into the BOADICEA breast cancer risk model. *Genet. Med.* **18** 1190–1198. https://doi.org/10.1038/gim.2016.31.

[43] MACMAHON, B. and COLE, P. (1969). Endocrinology and epidemiology of breast cancer. *Cancer* **24** 1146–1150. https://doi.org/10.1002/1097-0142(196912)24:6%3C1146::aid-cncr2820240612%3E3.0.co;2-0.

[44] MAVADDAT, N., MICHAILIDOU, K., DENNIS, J., LUSH, M., FACHAL, L., LEE, A., TYRER, J. P., CHEN, T.-H., WANG, Q. et al. (2019). Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am. J. Hum. Genet.* **104** 21–34. https://doi.org/10.1016/j.ajhg.2018.11.002.

[45] MCCORMACK, V. A. and DOS SANTOS SILVA, I. (2006). Breast density and parenchymal patterns as markers of breast cancer risk: A meta-analysis. *Cancer Epidemiol. Biomark. Prev.* **15** 1159–1169. https://doi.org/10.1158/1055-9965.epi-06-0034.

[46] MEALIFFE, M. E., STOKOWSKI, R. P., RHEES, B. K., PRENTICE, R. L., PETTINGER, M. and HINDS, D. A. (2010). Assess-

ment of clinical validity of a breast cancer risk model combining genetic and clinical information. *J. Natl. Cancer Inst.* **102** 1618–1627. https://doi.org/10.1093/jnci/djq388.

[47] COLLABORATIVE GROUP ON HORMONAL FACTORS IN BREAST CANCER (2012). Menarche, menopause, and breast cancer risk: Individual participant meta-analysis, including 118–964 women with breast cancer from 117 epidemiological studies. *Lancet Oncol.* **13** 1141–1151. https://doi.org/10.1016/s1470-2045(12)70425-4.

[48] PARMIGIANI, G., BERRY, D. A. and AGUILAR, O. (1998). Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2. *Am. J. Hum. Genet.* **62** 145–158. https://doi.org/10.1086/301670.

[49] PIKE, M. C., KRAILO, M. D., HENDERSON, B. E., CASAGRANDE, J. T. and HOEL, D. G. (1983). 'Hormonal' risk factors, 'breast tissue age' and the age-incidence of breast cancer. *Nature* **303** 767–770.

[50] PRENTICE, R. L., CAAN, B., CHLEBOWSKI, R. T., PATTERSON, R., KULLER, L. H., OCKENE, J. K., MARGOLIS, K. L., LIMACHER, M. C., MANSON, J. E. et al. (2006). Low-fat dietary pattern and risk of invasive breast cancer. *JAMA* **295** 629. https://doi.org/10.1001/jama.295.6.629.

[51] REEVES, G. K., BERAL, V., GREEN, J., GATHANI, T. and BULL, D. (2006). Hormonal therapy for menopause and breast-cancer risk by histological type: A cohort study and meta-analysis. *Lancet Oncol.* **7** 910–918. https://doi.org/10.1016/S1470-2045(06)70911-1.

[52] RICE, M. S., TWOROGER, S. S., HANKINSON, S. E., TAMIMI, R. M., ELIASSEN, A. H., WILLETT, W. C., COLDITZ, G. and ROSNER, B. (2017). Breast cancer risk prediction: An update to the Rosner–Colditz breast cancer incidence model. *Breast Cancer Res. Treat.* **166** 227–240. https://doi.org/10.1007/s10549-017-4391-5.

[53] RIGONI-STERN (1842). Fatti statistici relativi alle malattie cancerose. *Giorn. Prog. Patol. Terap.* **2** 507–517.

[54] ROCKHILL, B., SPIEGELMAN, D., BYRNE, C., HUNTER, D. J. and COLDITZ, G. A. (2001). Validation of the Gail et al. Model of breast cancer risk prediction and implications for chemoprevention. *J. Natl. Cancer Inst.* **93** 358–366. https://doi.org/10.1093/jnci/93.5.358.

[55] ROSNER, B. A., COLDITZ, G. A., HANKINSON, S. E., SULLIVAN-HALLEY, J., LACEY, J. V. and BERNSTEIN, L. (2013). Validation of Rosner–Colditz breast cancer incidence model using an independent data set, the California Teachers Study. *Breast Cancer Res. Treat.* **142** 187–202. https://doi.org/10.1007/s10549-013-2719-3.

[56] SASIENI, P. and BRENTNALL, A. R. (2017). On standardized relative survival. *Biometrics* **73** 473–482. MR3665964 https://doi.org/10.1111/biom.12578

[57] SASLOW, D., BOETES, C., BURKE, W., HARMS, S., LEACH, M. O., LEHMAN, C. D., MORRIS, E., PISANO, E., SCHNALL, M. et al. (2007). American cancer society guidelines for breast screening with MRI as an adjunct to mammography. *CA Cancer J. Clin.* **57** 75–89. https://doi.org/10.3322/canjclin.57.2.75.

[58] SCHOEMAKER, M. J., NICHOLS, H. B., WRIGHT, L. B., BROOK, M. N., JONES, M. E., O'BRIEN, K. M., ADAMI, H.-O., BAGLIETTO, L., BERNSTEIN, L. et al. (2018). Association of body mass index and age with subsequent breast cancer risk in premenopausal women. *JAMA Oncol.* **4** e181771. https://doi.org/10.1001/jamaoncol.2018.1771.

[59] STEEL, M., THOMPSON, A. and CLAYTON, J. (1991). Genetic aspects of breast cancer. *Br. Med. Bull.* **47** 504–518.

[60] TEAMS, F. C. (2010). Mammographic surveillance in women younger than 50 years who have a family history of breast cancer: Tumour characteristics and projected effect on mortality in the prospective, single-arm, FH01 study. *Lancet Oncol.* **11** 1127–1134. https://doi.org/10.1016/s1470-2045(10)70263-1.

[61] TICE, J. A., CUMMINGS, S. R., SMITH-BINDMAN, R., ICHIKAWA, L., BARLOW, W. E. and KERLIKOWSKE, K. (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** 337–347.

[62] TICE, J. A., MIGLIORETTI, D. L., LI, C.-S., VACHON, C. M., GARD, C. C. and KERLIKOWSKE, K. (2015). Breast density and benign breast disease: Risk assessment to identify women at high risk of breast cancer. *J. Clin. Oncol.* **33** JCO.2015.60.8869–3143. https://doi.org/10.1200/jco.2015.60.8869.

[63] TYRER, J., DUFFY, S. W. and CUZICK, J. (2004). A breast cancer prediction model incorporating familial and personal risk factors. *Stat. Med.* **23** 1111–1130. https://doi.org/10.1002/sim.1668.

[64] VACHON, C. M., PANKRATZ, V. S., SCOTT, C. G., HAEBERLE, L., ZIV, E., JENSEN, M. R., BRANDT, K. R., WHALEY, D. H., OLSON, J. E. et al. (2015). The Contributions of Breast Density and Common Genetic Variation to Breast Cancer Risk. *J Natl Cancer Inst* **107** dju397+. https://doi.org/10.1093/jnci/dju397.

[65] VAN VEEN, E. M., BRENTNALL, A. R., BYERS, H., HARKNESS, E. F., ASTLEY, S. M., SAMPSON, S., HOWELL, A., NEWMAN, W. G., CUZICK, J. et al. (2018). Use of single-nucleotide polymorphisms and mammographic density plus classic risk factors for breast cancer risk prediction. *JAMA Oncol.* **4** 476. https://doi.org/10.1001/jamaoncol.2017.4881.

[66] WACHOLDER, S., HARTGE, P., PRENTICE, R., GARCIA-CLOSAS, M., FEIGELSON, H. S., DIVER, W. R., THUN, M. J., COX, D. G., HANKINSON, S. E. et al. (2010). Performance of common genetic variants in breast-cancer risk models. *N. Engl. J. Med.* **362** 986–993. https://doi.org/10.1056/nejmoa0907727.

[67] WANG, C., BRENTNALL, A. R., CUZICK, J., HARKNESS, E. F., EVANS, D. G. and ASTLEY, S. (2017). A novel and fully automated mammographic texture analysis for risk prediction: Results from two case-control studies. *Breast Cancer Res.* **19** 114. https://doi.org/10.1186/s13058-017-0906-6.

[68] WARWICK, J., BIRKE, H., STONE, J., WARREN, R. M. L., PINNEY, E., BRENTNALL, A. R., DUFFY, S. W., HOWELL, A. and CUZICK, J. (2014). Mammographic breast density refines Tyrer–Cuzick estimates of breast cancer risk in high-risk women: findings from the placebo arm of the International Breast Cancer Intervention Study I. *Breast Cancer Research* **16** 451+. https://doi.org/10.1186/s13058-014-0451-5.

[69] WOLFE, J. N. (1976). Breast patterns as an index of risk for developing breast cancer. *Am. J. Roentgenol.* **126** 1130–1137.

[70] WOOD, S. N. (2006). *Generalized Additive Models: An Introduction with R. Texts in Statistical Science Series*. CRC Press/CRC, Boca Raton, FL. MR2206355

[71] WU, Y., ZHANG, D. and KANG, S. (2013). Physical activity and risk of breast cancer: A meta-analysis of prospective studies. *Breast Cancer Res. Treat.* **137** 869–882. https://doi.org/10.1007/s10549-012-2396-7.

[72] ZHANG, X., RICE, M., TWOROGER, S. S., ROSNER, B. A., ELIASSEN, A. H., TAMIMI, R. M., JOSHI, A. D., LINDSTROM, S., QIAN, J. et al. (2018). Addition of a polygenic risk score, mammographic density, and endogenous hormones to existing breast cancer risk prediction models: A nested case-control study. *PLoS Med.* **15** e1002644. https://doi.org/10.1371/journal.pmed.1002644.