Predicting 10-year breast cancer mortality risk in the general female population in England: a model development and validation study

Ash Kieran Clift, Gary S Collins, Simon Lord, Stavros Petrou, David Dodwell, Michael Brady, Julia Hippisley-Cox

Lancet Digit Health 2023;
5: e571–81

Cancer Research UK Oxford Centre (A K Clift MBBS), **Nuffield Department of Primary Care Health Sciences** (A K Clift, Prof S Petrou PhD, Prof J Hippisley-Cox MD), **Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences** (Prof G S Collins PhD), **Department of Oncology** (S Lord DPhil, Prof M Brady PhD), **and Nuffield Department of Population Health** (D Dodwell MD), **University of Oxford, UK**

Correspondence to:
Dr Ash Kieran Clift, Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford OX2 6GG, UK
ashley.clift@phc.ox.ac.uk

Summary

Background Identifying female individuals at highest risk of developing life-threatening breast cancers could inform novel stratified early detection and prevention strategies to reduce breast cancer mortality, rather than only considering cancer incidence. We aimed to develop a prognostic model that accurately predicts the 10-year risk of breast cancer mortality in female individuals without breast cancer at baseline.

Methods In this model development and validation study, we used an open cohort study from the QResearch primary care database, which was linked to secondary care and national cancer and mortality registers in England, UK. The data extracted were from female individuals aged 20–90 years without previous breast cancer or ductal carcinoma in situ who entered the cohort between Jan 1, 2000, and Dec 31, 2020. The primary outcome was breast cancer-related death, which was assessed in the full dataset. Cox proportional hazards, competing risks regression, XGBoost, and neural network modelling approaches were used to predict the risk of breast cancer death within 10 years using routinely collected health-care data. Death due to causes other than breast cancer was the competing risk. Internal–external validation was used to evaluate prognostic model performance (using Harrell's C, calibration slope, and calibration in the large), performance heterogeneity, and transportability. Internal–external validation involved dataset partitioning by time period and geographical region. Decision curve analysis was used to assess clinical utility.

Findings We identified data for 11 626 969 female individuals, with 70 095 574 person-years of follow-up. There were 142 712 (1·2%) diagnoses of breast cancer, 24 043 (0·2%) breast cancer-related deaths, and 696 106 (6·0%) deaths from other causes. Meta-analysis pooled estimates of Harrell's C were highest for the competing risks model (0·932, 95% CI 0·917–0·946). The competing risks model was well calibrated overall (slope 1·011, 95% CI 0·978–1·044), and across different ethnic groups. Decision curve analysis suggested favourable clinical utility across all age groups. The XGBoost and neural network models had variable performance across age and ethnic groups.

Interpretation A model that predicts the combined risk of developing and then dying from breast cancer at the population level could inform stratified screening or chemoprevention strategies. Further evaluation of the competing risks model should comprise effect and health economic assessment of model-informed strategies.

Funding Cancer Research UK.

Introduction

Screening mammography and improvements in treatments have reduced breast cancer mortality over recent decades. Further reducing the public health and societal burdens of breast cancer[1] could be achievable with stratified screening or prevention strategies structured around accurate estimation of individual risks.[2,3] Risk-stratified early detection could tailor screening intensity, starting ages, or modality,[2,3] and chemoprevention programmes could use cost-effective agents such as tamoxifen[4] or anastrozole.[5] These approaches are typically envisioned to be guided by risk-stratification methods that predict absolute or relative risks of breast cancer diagnosis.[6]

However, predicted risks of breast cancer incidence can correlate poorly or inversely with risks of mortality, screening mammography is associated with overdiagnosis,[7–9] and tumour subtypes vary in aggressiveness.[10] Furthermore, chemoprevention has differential effects on disease subtypes, and whether these therapies reduce breast cancer mortality is uncertain.[4,5] Identification of female individuals at the greatest risk of developing life-threatening cancers could inform new approaches to early detection and prevention by directing them towards additional interventions based on their mortality risk,[11] rather than on the basis of their risk of breast cancer diagnosis.[12] Models for predicting breast cancer mortality in female individuals without breast cancer at baseline do not yet exist.

This study sought to develop a prognostic model that accurately predicts the 10-year risk of breast cancer mortality in females without breast cancer at baseline using a national, population-representative, linked electronic health-care record dataset of over 11·6 million female individuals.

Research in context

Evidence before this study

We searched PubMed with the terms: ("breast") AND ("mammography" OR "screening" OR "prevention") AND ("risk#adapted" OR "risk#stratified" OR "personalised" OR "personalized" OR "tailored" OR "risk#based") on Nov 1, 2020, then updated this search on June 1, 2022. Papers published in English that developed or validated clinical prediction models were identified. Several models predicting risk of incident breast cancer exist, such as IBIS, BOADICEA, BRCAPRO, BCRAT, the BCSC model, and QCancer (Breast), with varying extents of validation. There is evidence of screening-related overdiagnosis, uncertainty regarding the effects of chemoprevention agents in reducing breast cancer death, and variable or inverse relationships between predicted incident cancer risk and outcomes after diagnosis. Therefore, predicting mortality rather than incidence could be beneficial for stratifying breast cancer early detection or prevention. However, no models were found that estimate the risk of breast cancer mortality in women without breast cancer.

Added value of this study

To our knowledge, this regression and machine learning modelling study of 11·6 million female individual's linked electronic health records is the largest yet undertaken to develop and evaluate clinical prediction models in breast cancer. The study is also the first to develop a model to estimate the risk of breast cancer mortality at the population level. Internal–external validation enabled the robust comparison of results from several modelling methods. Competing risks regression yielded a model that discriminated well, was well calibrated, and was associated with favourable net benefit across all age groups examined (as assessed by decision curve analysis).

Implications of all the available evidence

We present a new model that directly estimates the risk of 10-year breast cancer mortality in the general population of females without breast cancer. Pending further evaluation, such as external validation, randomised impact studies, and cost-effectiveness analyses, the final model could aid in the identification of women at high risk who are currently too young to be screened, or women suitable for chemoprevention regardless of age.

Methods

Study design

In this model development and validation study, we explored two regression (ie, Cox proportional hazards and competing risks), and two machine learning approaches (ie, XGBoost and a feed-forward neural network) to predict the 10-year risk of breast cancer death in females without breast cancer. Models were evaluated using an internal–external validation framework, which we previously applied to develop and compare models predicting 10-year risk of breast cancer mortality in women with invasive breast cancer.[13]

See Online for appendix

Data were obtained from the QResearch database between Jan 1, 2000, and Dec 31, 2020. QResearch covers more than 1500 general practices in England, UK, with individual-level linkage across primary care, National Health Service (NHS) Digital's Hospital Episode Statistics (HES), the national cancer registry, and the Office for National Statistics mortality register.

The study received approval from the QResearch Scientific committee (reference OX129). Ethical approval for the QResearch database is with the Derby Research Ethics Committee (18/EM/0400). The protocol for this study has been published.[14] The TRIPOD statement was used for study reporting.[15]

Participants

This study targeted the adult female population aged 20 years and older. An open cohort of females (ie, self-reported female sex) aged 20–90 years at entry into the QResearch database was used. The cohort data spanned Jan 1, 2000, to Dec 31, 2020. We obtained data on mortality and cause of death from death certificates, which are completed by a clinician who cared for the individual before death. Individuals entered the cohort from the latest of three events: their 20th birthday, 1 year from the date of general practice registration, or 1 year from the date of the practice contributing data to QResearch. Cohort participants with recorded existing or previous diagnoses of invasive breast carcinoma or ductal carcinoma in situ on general practice, HES, or cancer registry records were excluded (appendix p 22).

Individual informed consent was not required for this analysis of anonymised health records, and individuals could opt out of clinical data sharing.

Outcomes and candidate predictors

The primary outcome was breast cancer mortality. We defined breast cancer mortality as breast cancer being recorded as a primary or contributory cause of death on ONS death certificates. This definition ascertained direct deaths and deaths indirectly related to the malignancy. Follow-up was from cohort entry to date of breast cancer-related death or censoring (ie, alive at study end date; left the general practice or cohort; or died from another, non-breast cancer-related cause [ie, competing event]). We identified female individuals with incident breast cancer diagnoses, defined as presence of Systemized Nomenclature of Medicine or Read codes in the general practice record and International Classification of Diseases-10 codes in HES or the cancer registry.

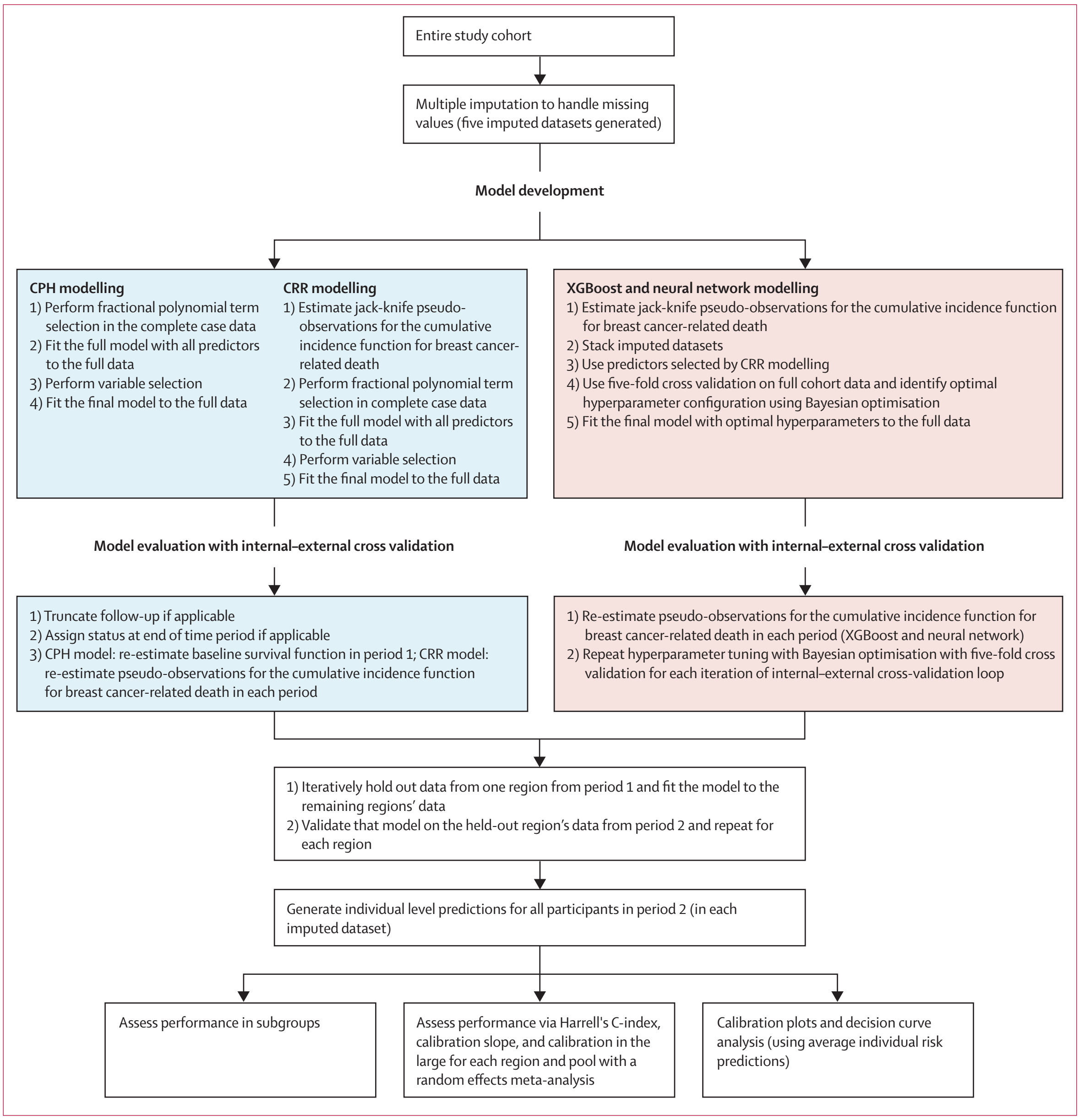
Entire study cohort
Multiple imputation to handle missing values (five imputed datasets generated)
Model development
CPH modelling
1) Perform fractional polynomial term selection in the complete case data
2) Fit the full model with all predictors to the full data
3) Perform variable selection
4) Fit the final model to the full data
CRR modelling
1) Estimate jack-knife pseudo-observations for the cumulative incidence function for breast cancer-related death
2) Perform fractional polynomial term selection in complete case data
3) Fit the full model with all predictors to the full data
4) Perform variable selection
5) Fit the final model to the full data
XGBoost and neural network modelling
1) Estimate jack-knife pseudo-observations for the cumulative incidence function for breast cancer-related death
2) Stack imputed datasets
3) Use predictors selected by CRR modelling
4) Use five-fold cross validation on full cohort data and identify optimal hyperparameter configuration using Bayesian optimisation
5) Fit the final model with optimal hyperparameters to the full data
Model evaluation with internal–external cross validation
Model evaluation with internal–external cross validation
1) Truncate follow-up if applicable
2) Assign status at end of time period if applicable
3) CPH model: re-estimate baseline survival function in period 1; CRR model: re-estimate pseudo-observations for the cumulative incidence function for breast cancer-related death in each period
1) Re-estimate pseudo-observations for the cumulative incidence function for breast cancer-related death in each period (XGBoost and neural network)
2) Repeat hyperparameter tuning with Bayesian optimisation with five-fold cross validation for each iteration of internal–external cross-validation loop
1) Iteratively hold out data from one region from period 1 and fit the model to the remaining regions' data
2) Validate that model on the held-out region's data from period 2 and repeat for each region
Generate individual level predictions for all participants in period 2 (in each imputed dataset)
Assess performance in subgroups
Assess performance via Harrell's C-index, calibration slope, and calibration in the large for each region and pool with a random effects meta-analysis
Calibration plots and decision curve analysis (using average individual risk predictions)

Figure 1: **Model development and internal–external cross validation**
Models were internally–externally cross validated. Non-randomly splitting the data into non-overlapping, structurally different units (ie, by geographical region or time period), is a form of external validation. When there are multiple non-random splits, this process can be iterated so that each structurally different unit is held out to validate a model developed on all the other data. The predictions generated for each individual when held out can be used to assess model performance. This approach can internally–externally validate a model that is fitted to the entire study dataset.[17] This approach not only assesses model performance, but also incorporates an evaluation of the model's transportability to new settings by simulating the same process in the derivation dataset. CPH=Cox proportional hazards. CRR=competing risks regression.

We identified candidate predictor variables associated with the risk of breast cancer diagnosis or mortality in clinical or epidemiological literature. Candidate predictor values at cohort entry, or most recently recorded before entry (without time restrictions), were used. Medication use was defined as ever receiving three or more prescriptions for the drug. Fractional polynomials (a maximum of two powers) were explored for age, BMI, and Townsend deprivation score for the regression models using participants with complete data. The powers selected were those that minimised the deviance. This analysis was done separately for the regression models to permit different predictor–outcome associations and used the mfp command. Two-way interaction terms between age and family history of breast cancer were explored in the regression models (ie, continuous by categorical).[12]

	Overall study cohort (n=11 626 969)	Period 1 sub-cohort (n=6 151 399)	Period 2 sub-cohort (n= 5 475 570)
Age at cohort entry	41·78 (18·13)	42·95 (18·44)	40·48 (17·67)
BMI at entry			
Mean (SD)	25·37 (5·46)	25·02 (5·09)	25·68 (5·74)
Not recorded	4 967 520 (42·7%)	3 042 901 (49·5%)	1 924 619 (35·2%)
Townsend deprivation score			
Mean (SD)	0·71 (3·23)	0·50 (3·24)	0·94 (3·21)
Not recorded	50 889 (0·4%)	18 912 (0·3%)	31 977 (0·6%)
Ethnic group*			
Black	445 720 (3·8%)	190 088 (3·1%)	255 632 (4·7%)
Chinese	132 583 (1·1%)	39 643 (0·6%)	92 940 (1·7%)
Other Asian	189 635 (1·6%)	66 538 (1·1%)	123 097 (2·3%)
South Asian	507 829 (4·4%)	199 903 (3·3%)	307 926 (5·6%)
White	6 168 419 (53·1%)	2 751 371 (44·7%)	3 417 048 (62·4%)
Other	328 396 (2·8%)	101 147 (1·6%)	227 249 (4·2%)
Not recorded	3 854 387 (33·2%)	2 802 709 (45·6%)	1 051 678 (19·2%)
Smoking status			
Non-smoker	5 045 101 (43·4%)	2 207 095 (35·9%)	2 838 006 (51·8%)
Ex-smoker	1 445 584 (12·4%)	655 998 (10·7%)	789 586 (14·4%)
Light smoker (1–9 cigarettes per day)	1 318 132 (11·3%)	688 109 (11·2%)	630 023 (11·5%)
Moderate smoker (10–19 cigarettes per day)	308 372 (2·7%)	173 942 (2·8%)	134 430 (2·5%)
Heavy smoker (≥20 cigarettes per day)	133 108 (1·1%)	86 337 (1·4%)	46 771 (0·9%)
Not recorded	3 376 672 (29·0%)	2 339 918 (38·0%)	1 036 754 (18·9%)
Alcohol intake			
Non-drinker	4 120 142 (35·4%)	1 994 174 (32·4%)	2 125 968 (38·8%)
Minimal (<1 unit per day)	1 580 548 (13·6%)	770 767 (12·5%)	809 781 (14·8%)
Light (1–2 units per day)	601 071 (5·2%)	289 449 (4·7%)	311 622 (5·7%)
Moderate (3–6 units per day)	246 366 (2·1%)	128 708 (2·1%)	117 658 (2·2%)
Heavy (7–9 units per day)	9823 (0·1%)	3865 (0·1%)	5958 (0·1%)
Very heavy (>9 units per day)	17 467 (0·2%)	2510 (<0·1%)	14 957 (0·3%)
Not recorded	5 051 552 (43·5%)	2 961 926 (48·2%)	2 089 626 (38·2%)
Benign breast disease	282 663 (2·4%)	142 108 (2·3%)	140 555 (2·6%)
Endometriosis	151 158 (1·3%)	59 717 (1·0%)	91 441 (1·7%)
Polycystic ovarian syndrome	197 886 (1·7%)	57 951 (0·9%)	139 935 (2·6%)
Hysterectomy	99 439 (0·9%)	31 051 (0·5%)	68 388 (1·3%)
Previous gynaecological cancer	26 626 (0·2%)	11 264 (0·2%)	15 362 (0·3%)

(Table 1 continues on next page)

Code lists used to define predictors and outcomes are available online.[16]

Procedures for missing data

Before model development, data for the entire study cohort was multiply imputed to handle missing data for alcohol intake, smoking status, BMI, Townsend deprivation score, and ethnicity (figure 1). Multiple imputation with chained equations (five imputations for computational considerations) was used under the missing at random assumption.[18] The imputation model included all candidate predictor parameters, the endpoint indicator, the Nelson-Aalen cumulative hazard estimate,[18] and decade of cohort entry (ie, period 1 [Jan 1, 2000, to Dec 31, 2009] or period 2 [Jan 1, 2010, to Dec 31, 2020]). Natural logarithms of BMI values were imputed for normality, then exponentiated after imputation for analysis. Multiply imputed data were used throughout all model fitting and evaluation steps.

Modelling strategy

Each model was fit to the entire cohort, and their respective performance was estimated using internal–external cross validation[17] involving non-random splitting by decade of entry (ie, period 1 and period 2) and geographical region (as categorised in the QResearch database ie, East Midlands, East of England, London, North East, North West, South Central, South East, South West, West Midlands, and Yorkshire and the Humber). There was no patient overlap between periods. Prediction models are developed using currently available data but implemented prospectively—in the context of varying baseline rates and case mix, performance might change over time on implementation in new settings. Splitting the dataset into structurally distinct subunits in terms of location and time is a form of external validation. Internal–external validation can estimate how well a model might be generalisable to temporally or geographically different settings by simulating the same process, (ie, developing a model in one sample and applying it to a later, distinct sample). This method is a more informative evaluation than assessing generalisation to one randomly partitioned subset of data with similar characteristics.[17] Models were iteratively refit to all but one region in period 1 and their performance estimated in the held-out region's data from period 2 (figure 1).

For the Cox proportional hazards model, a full model was fit incorporating all candidate predictors. Predictors associated with an exponentiated coefficient (ie, hazard ratio [HR]) greater than 1·1 or less than 0·9, and with p<0·01, were selected. The final model with these predictors was then fitted to the entire cohort. Model coefficients, SEs, and baseline survival functions (with continuous covariates centred at their means and binary variables assigned a value of 0) were estimated in and pooled across imputations using Rubin's rules.[19]

The competing risks regression model was developed using jack-knife pseudo-observations for the Aalen-Johansen cumulative incidence function at 10 years estimated in the full cohort, regressed on the candidate predictors using a generalised linear model with a complementary log-log link function and robust SEs.[20] The exponentiated coefficients of this model are interpretable as subdistribution HRs. The same predictor selection criteria were used as for the Cox proportional hazards model, as were Rubin's rules to combine results across imputations.

For benchmarking, predictors selected by the competing risks regression were included in the machine learning models. The XGBoost and neural network models used the pseudo-observations as a continuous outcome variable,[21] enabling handling of censored, time-to-event data in a competing risks setting. Categorical predictors were handled as dummy variables. Continuous predictors were left unscaled for XGBoost (a tree-based approach that handles unscaled data), but minimum–maximum scaled for the neural networks as these models are more affected by variable scale than XGBoost. Imputations were stacked, forming a single long dataset so that all imputations were used for model fitting. A feed-forward neural network architecture was selected due to the tabular nature of the data and its low dimensionality. Feed forward refers to the connections between nodes in each layer not forming cycles—information flows though the network unidirectionally.

The root mean squared error between predicted and observed pseudo-observations was used as the loss function,[21] due to the continuous target variable. 5-fold cross-validation with Bayesian optimisation (using the expected improvement acquisition function) was used on the whole dataset for hyperparameter tuning to identify the optimal configurations to minimise the root mean squared error. Model architectures, the hyperparameters tuned, ranges, and final configurations are available in the appendix (p 6).

Performance assessment with internal–external cross validation

Model performance was assessed using the predictions generated during the IECV process (figure 1).[17]

Predicted risks from the Cox model were calculated by combining the baseline survival function at 10 years with the linear predictor. For the competing risks regression model, the following transformation of the linear predictor (including constant) was used to calculate the predicted event probability: $1-\exp(-\exp(X\beta))$. Machine learning model hyperparameter tuning with Bayesian optimisation was recapitulated in IECV. This nested cross-validation strategy avoided using the same data concomitantly for tuning and evaluation.

Calibration slope, calibration in the large, and Harrell's C-index were estimated in each geographical region and pooled using random effects meta-analysis with the Hartung-Knapp-Sidik-Jonkmann method[22] to calculate point estimates, 95% CIs, and 95% prediction intervals. Calibration slope describes the spread of predicted risks and whether they are too extreme, and the ideal value is 1. Calibration in the large measures whether a model systematically overpredicts or underpredicts risk, and the ideal value is 0. Harrell's C is a discrimination metric in the range of 0·5–1, with a value of 1 meaning perfect separation between individuals who have the event and those who did not. Harrell's C was weighted by inverse probability of censoring for the competing risks regression and machine learning models. Royston and Sauerbrei's R^2 and D statistics were calculated for the Cox proportional hazards model.[23] To understand the heterogeneity of model performance across societally relevant groups,

	Overall study cohort (n=11 626 969)	Period 1 sub-cohort (n=6 151 399)	Period 2 sub-cohort (n= 5 475 570)
(Continued from previous page)			
Oral contraceptive pill use (ever)	1 372 633 (11·8%)	519 506 (8·5%)	853 127 (15·6%)
Duration of recent† oestrogen-only HRT			
None	11 467 510 (98·6%)	6 037 813 (98·2%)	5 429 697 (99·2%)
<1 year	58 156 (0·5%)	40 563 (0·7%)	17 593 (0·3%)
1–2·9 years	34 566 (0·3%)	27 071 (0·4%)	7495 (0·1%)
3–4·9 years	25 760 (0·2%)	21 024 (0·3%)	4736 (0·1%)
5–9·9 years	30 254 (0·3%)	22 943 (0·4%)	7311 (0·1%)
≥10 years	10 723 (0·1%)	1985 (<0·1%)	8738 (0·2%)
Duration of past‡ oestrogen-only HRT			
None	11 551 573 (99·4%)	6 134 698 (99·7%)	5 416 875 (98·9%)
<1 year	35 352 (0·3%)	11 035 (0·2%)	24 317 (0·4%)
1–2·9 years	12 685 (0·1%)	3297 (0·1%)	9388 (0·2%)
3–4·9 years	8732 (0·1%)	1318 (<0·1%)	7414 (0·1%)
5–9·9 years	13 071 (0·1%)	956 (<0·1%)	12 115 (0·2%)
≥10 years	5 556 (0·1%)	95 (<0·1%)	5461 (0·1%)
Duration of recent† combined HRT			
None	11 339 053 (97·52%)	5 929 384 (96·4%)	5 409 669 (98·8%)
<1 year	85 664 (0·7%)	65 937 (1·1%)	19 727 (0·4%)
1–2·9 years	68 532 (0·6%)	56 284 (0·9%)	12 248 (0·2%)
3–4·9 years	52 565 (0·5%)	44 811 (0·7%)	7754 (0·1%)
5–9·9 years	63 127 (0·5%)	50 867 (0·8%)	12 260 (0·2%)
≥10 years	18 028 (0·2%)	4116 (0·1%)	13 912 (0·3%)
Duration of past‡ combined HRT			
None	11 489 012 (98·8%)	6 124 006 (99·6%)	5 365 006 (98·0%)
<1 year	48 225 (0·4%)	15 520 (0·3%)	32 705 (0·6%)
1–2·9 years	26 529 (0·2%)	6177 (0·1%)	20 352 (0·4%)
3–4·9 years	20 172 (0·2%)	2983 (0·1%)	17 189 (0·3%)
5–9·9 years	31 508 (0·3%)	2372 (<0·1%)	29 136 (0·5%)
≥10 years	11 523 (0·1%)	341 (<0·1%)	11 182 (0·2%)
Family history of breast cancer	177 368 (1·5%)	64 718 (1·1%)	112 650 (2·1%)
Family history of gynaecological cancer	36 932 (0·3%)	12 349 (0·2%)	24 583 (0·5%)
Previous lung cancer	9414 (0·1%)	3529 (0·1%)	885 (0·1%)
Previous haematological cancer	31 637 (0·3%)	13 050 (0·2%)	18 587 (0·3%)
Previous thyroid cancer	6009 (0·1%)	1981 (<0·1%)	4028 (0·1%)
Type 1 diabetes	20 479 (0·2%)	9782 (0·2%)	10 697 (0·2%)
Type 2 diabetes	311 725 (2·7%)	129 256 (2·1%)	182 469 (3·3%)
Chronic kidney disease			
None or stage 2	11 471 953 (98·7%)	6 129 623 (99·7%)	5 342 330 (97·6%)
Stage 3	136 585 (1·2%)	15 077 (0·3%)	121 508 (2·2%)
Stage 4	7972 (0·1%)	1113 (<0·1%)	6859 (0·1%)
Stage 5 (including end-stage renal failure and dialysis)	10 459 (0·1%)	5586 (0·1%)	4873 (0·1%)
Hypertension	1 073 831 (9·2%)	524 255 (8·5%)	549 576 (10·0%)
Ischaemic heart disease	255 299 (2·2%)	147 567 (2·4%)	107 732 (2·0%)
Chronic liver disease	30 550 (0·3%)	10 076 (0·2%)	20 474 (0·4%)
Systemic lupus erythematosus	14 541 (0·1%)	6483 (0·1%)	8058 (0·2%)
Vasculitis	63 329 (0·5%)	28 416 (0·5%)	34 913 (0·6%)
Psychotic condition	87 334 (0·8%)	38 072 (0·6%)	49 262 (0·9%)
Anti-psychotic medication use (ever)	119 285 (1·0%)	54 964 (0·9%)	64 321 (1·2%)

(Table 1 continues on next page)

	Overall study cohort (n=11 626 969)	Period 1 sub-cohort (n=6 151 399)	Period 2 sub-cohort (n= 5 475 570)
(Continued from previous page)			
Thiazide use (ever)	433 488 (3·7%)	233 350 (3·8%)	200 138 (3·7%)
β-blocker use (ever)	547 073 (4·7%)	272 165 (4·4%)	274 908 (5·0%)
Renin-angiotensin-aldosterone axis inhibitor use (ever)	522 320 (4·5%)	195 058 (3·2%)	327 262 (6·0%)
Angiotensin converting enzyme inhibitor use (ever)	441 264 (3·8%)	174 109 (2·8%)	267 155 (4·9%)
Calcium channel blocker use (ever)	392 027 (3·4%)	141 830 (2·3%)	250 197 (4·6%)
Tricyclic antidepressant use (ever)	520 745 (4·5%)	242 492 (3·9%)	278 253 (5·1%)
Monoamine oxidase inhibitor use (ever)	3328 (<0·1%)	2277 (<0·1%)	1051 (<0·1%)
Selective serotonin reuptake inhibitor use (ever)	886 309 (7·6%)	263 703 (4·3%)	622 606 (11·4%)

Data are n (%) and mean (SD) unless otherwise specified. Medication use is defined as the receipt of three or more prescriptions in the primary care records. HRT=hormone replacement therapy. *Ethnicity was self-reported and based on the UK Office for National Statistics classifications; as ethnic group was a variable of interest for examining performance heterogeneity, and event rates were low in some ethnic groups, some groups were collated: South Asian comprises Indian, Pakistani and Bangladeshi; Other Asian comprises non-Chinese Asian groups; and Black comprises Black Caribbean, Black African, and Other Black. †Recent refers to a duration of less than 5 years since last HRT prescription and cohort entry date. ‡Past refers to 5 or more years since last HRT prescription and the cohort entry date.

Table 1: **Baseline characteristics**

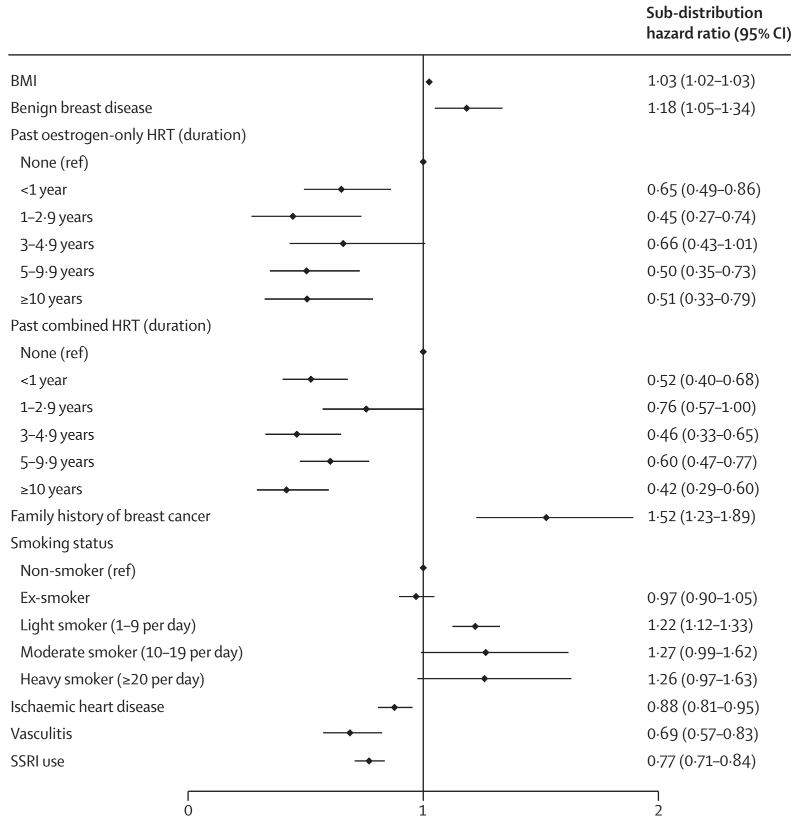

Figure 2: **10-year risk of breast cancer mortality in the competing risks regression model**
Forest plot showing the final competing risks regression model as its exponentiated coefficients with 95% CIs. Fractional polynomial terms for age and the constant term are not included due to scaling. SSRI=selective serotonin reuptake inhibitor.

IECV-derived predictions were used to estimate ethnic group-specific and age group-specific performance metrics. Age groups were informed by UK breast screening policy: pre-screening (20–49 years), screening (50–70 years), and post-screening (>70 years). Random effects meta-regression estimated the proportion of inter-regional heterogeneity in model performance accounted for by geographical differences in age, BMI, deprivation, and ethnic diversity (percentage of non-White individuals).

Smoothed calibration plots visualised model calibration across the predicted risks spectrum (stpsurv, stpci, and running commands in Stata).[24] Decision curve analysis compared the net benefit of each model overall and of each model by age group, accounting for the competing risk of death from other causes. Model sensitivity was assessed by the proportions of breast cancer deaths captured within the different cutoffs of predicted risk distributions.

Using statistics from Cancer Research UK,[25] assuming a mean follow-up of 6 years, 100 candidate predictor parameters (to permit all candidate predictors, plus interactions and transformations), a Cox-Snell R^2 of 0·0045 (15% of maximum Cox-Snell R^2=0·03 permitted by the underlying equations), and an age-standardised annual breast cancer mortality rate of 0·000334 (per person), the minimum sample size for regression model development was 199 500, with 400 outcome events (four events per predictor parameter).[26]

Analyses were done using Stata version 17 and R version 3.7.

Role of the funding source

The funder of the study had no role in the study design, data collection, data analysis, data interpretation, or the writing of this report.

Results

After excluding female individuals with a recorded history of previous or current breast cancer (n=152 870) or ductal carcinoma in situ diagnoses (n=5409), the final study cohort comprised 11 626 969 females (table 1, appendix p 22). During the 70 095 574 person-years of follow-up, there were 142 712 (1·2%) breast cancer diagnoses, 24 043 (0·2%) breast cancer-related deaths, and 696 106 (6·0%) deaths due to other causes. Median follow-up from cohort entry was 3·74 years (range 0·003–20·60), with a mean of 6·03 years (SD 5·90). After restricting to 10 years follow-up (ie, prediction horizon), there were 13 062 (0·1%) breast cancer-related deaths within 55 104 482 person-years (crude mortality rate 2·37 [95% CI 2·33–2·41] per 10 000 person-years).

In the temporally distinct subcohorts, there were 7999 breast cancer deaths in period 1 (crude mortality rate 2·66 [2·60–2·72] per 10 000 person-years), and 2712 in period 2 (1·54 [1·49–1·60] per 10 000 person-years). Ethnic group-specific and region-specific mortality rates are summarised in the appendix (pp 1–3).

	Cox proportional hazards model		Competing risks model		XGBoost		Neural network	
	Estimate (95% CI)	95% prediction interval	Estimate (95% CI)	95% prediction interval	Estimate (95% CI)	95% prediction interval	Estimate (95% CI)	95% prediction interval
Harrell's C index	0·854 (0·842 to 0·865)	0·822 to 0·885	0·932 (0·917 to 0·946)	0·886 to 0·977	0·839 (0·805 to 0·873)	0·737 to 0·942	0·771 (0·751 to 0·792)	0·718 to 0·792
Calibration slope	1·091 (0·991 to 1·191)	0·787 to 1·395	1·011 (0·978 to 1·044)	0·913 to 1·110	1·021 (0·989 to 1·052)	0·926 to 1·116	1·037 (1·003 to 1·071)	0·935 to 1·140
Calibration in the large	0·091 (–0·009 to 0·191)	–0·213 to 0·395	0·011 (–0·022 to 0·044)	–0·087 to 0·110	0·021 (–0·011 to 0·052)	–0·074 to 0·116	0·037 (0·003 to 0·071)	–0·065 to 0·140
Royston and Sauerbrei's D Statistic	2·397 (2·288 to 2·506)	2·117 to 2·677	..	..	..	..	..	..
Royston and Sauerbrei's R^2	0·579 (0·557 to 0·601)	0·523 to 0·636	..	..	..	..	..	..
Brier score	0·003 (0·002 to 0·003)	0·001 to 0·004	..	..	..	..	..	..

Royston and Sauerbrei's D Statistic and R^2 and Brier score are not estimable for the competing risks regression and machine learning models. Harrell's C ranges between 0·5 and 1, with the ideal value being 1 (denoting perfect discrimination between individuals who had the event and those who did not). Calibration slope measures whether or not risk predictions are too extreme or too moderate and has an ideal value of 1. Calibration in the large measures whether a model systematically overpredicts or underpredicts risk and has an ideal value of 0. The D statistic is a measure of discrimination and can be interpreted as the hazard ratio when the sample is split at the median of predicted risk. The R^2 is a measure of the variation in the time to event explained by a model (eg, 0·5 means 50%). The Brier score is a measure of prediction accuracy, which is the mean squared error between predictions and outcomes, and lower values are better.

Table 2: **Performance metrics for the four developed models**

Non-linear fractional polynomial terms were selected for age (–2 and 3) and BMI (–2 and –2) for the Cox model, and for age (1 and 2) for the competing risks regression model (appendix pp 16–17). During analysis, ethnicity was selected for inclusion in both regression models as coefficients for most non-White ethnic groups met the predictor selection criteria. For some ethnic groups, these coefficients were less than 0. As the envisioned use cases for these models include risk-based screening or prevention, including ethnicity as a predictor could influence eligibility. Modelling was repeated without ethnicity as a predictor, but ethnicity was considered when assessing performance heterogeneity. The final Cox and competing risks regression models (without ethnicity) are presented in the appendix (p 18) and figure 2, and in full (as their coefficients) in the appendix (pp 4–5, 18). The models including ethnicity are also available in the appendix (pp 19–20). The final Cox model comprised 13 predictors: age (two fractional polynomial terms), BMI (two fractional polynomial terms), benign breast disease, previous lung cancer, previous haematological cancer, smoking status, type 1 diabetes, type 2 diabetes, chronic kidney disease, chronic liver disease, vasculitis, psychotic condition, and anti-psychotic medication use.

The final competing risks regression model comprised 11 predictors: age (two fractional polynomial terms), BMI, past use of oestrogen-only hormone replacement therapy, past use of combined hormone replacement therapy, family history of breast cancer, smoking status, ischaemic heart disease, vasculitis, selective serotonin reuptake inhibitor use, and the interactions between fractional polynomial terms for age and family history of breast cancer.

The competing risks regression model had the highest discrimination, with a Harrell's C-index of 0·932 (95% CI 0·917–0·946; 95% prediction interval 0·886–0·977), whereas the neural network model had the lowest (Harrell's C statistic 0·771, 0·751–0·792; 0·718–0·792; table 2). The Cox, competing risks, and XGBoost models did not have any discernible miscalibration on summary measures, but the neural network did, albeit to a small extent.

On calibration plots (figure 3), all models tended towards overestimation at the very highest range of the predicted risk spectrum; miscalibration of the Cox model began at a lower range than the other models. All models also tended towards overestimation in individuals with the very highest predicted risks (eg, >2% 10-year risk); for the competing risks model, the tendency towards overestimation appeared to occur above a risk threshold of 0·015, which represents 0·7% of individuals.

The regression models generally discriminated well in ethnic subgroups (appendix p 8), although some CIs were wide. With the caveat of small event numbers, these models were generally well calibrated in most ethnic groups, apart from some miscalibration with the competing risks model in the Other Asian subgroup (slope 1·252, 95% CI 1·075–1·428). The XGBoost and neural network approaches had more inconsistent performance across different ethnic groups (appendix p 8), such as poor discrimination of both models in female individuals who were Black (Harrell's C for XGBoost: 0·569, 95% CI 0·418–0·720; for neural network: 0·623, 0·469–0·776). These results are compared with Harrell's C results of 0·863 (95% CI 0·847–0·880) for XGBoost and 0·788 (0·767–0·809) for neural network models in female individuals who were White (appendix p8).

More complex patterns of performance were observed across the age subgroups (appendix p 7). Although the Cox

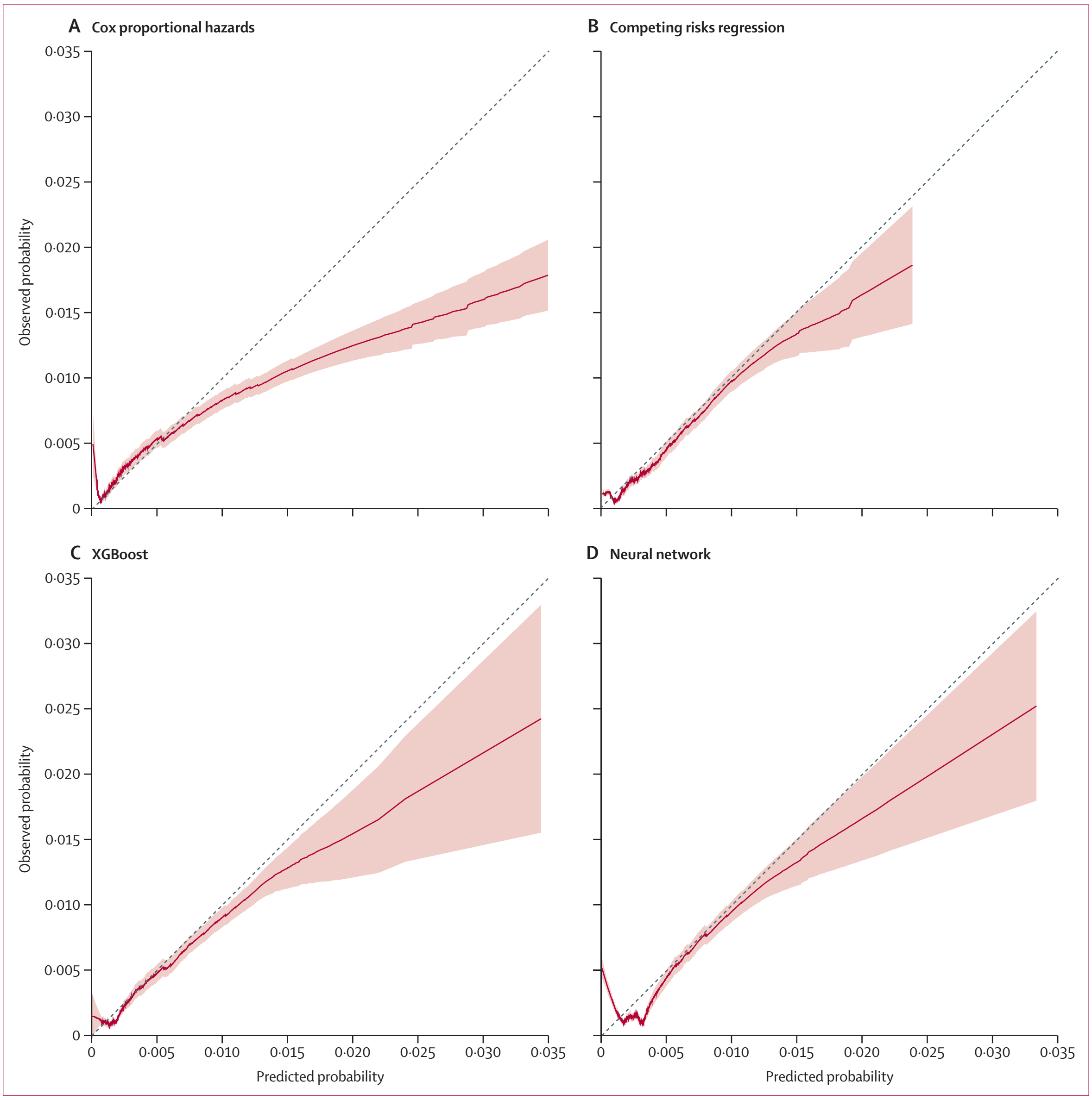

Figure 3: **Smoothed calibration plots of the alignment between predicted and observed risks for each of the models**
The red lines correspond to the calibration curve, with the red shaded area corresponding to the 95% CI. The dotted lines are a reference for perfect calibration (ie, perfect alignment between the observed and predicted probabilities).

model generally discriminated better across age groups than the XGBoost and neural network models, it was miscalibrated in the pre-screening age group (slope 1·771, 95% CI 1·558 to 1·954) and the post-screening age subgroup (0·120, –0·108 to 0·349). Discrimination of machine learning models in the pre-screening age group was poor (eg, Harrell's C for XGBoost 0·404, 95% CI 0·359 to 0·449 [appendix p 7]). The competing risks regression model did not show any miscalibration in any age group; discrimination was lower than overall metrics from IECV (table 2) due to the restricted age range (appendix p 7) but was highest of all models in the pre-screening and post-screening age subgroups.

The top 1% of predicted risks from each model captured at least 8% of all breast cancer deaths, and the highest 10% of predicted risks from each model captured at least 49% of all breast cancer deaths, suggesting potential for population stratification (appendix p 15).

Decision curve analysis (figure 4) showed that the neural network model was associated with the lowest net benefit compared with the other models. All other models had a similar or better net benefit association compared with the clinically unrealistic screen or treat all strategy. The regression models were associated with improved net benefit in individuals in the pre-screening age subgroup, and the competing risks model was associated

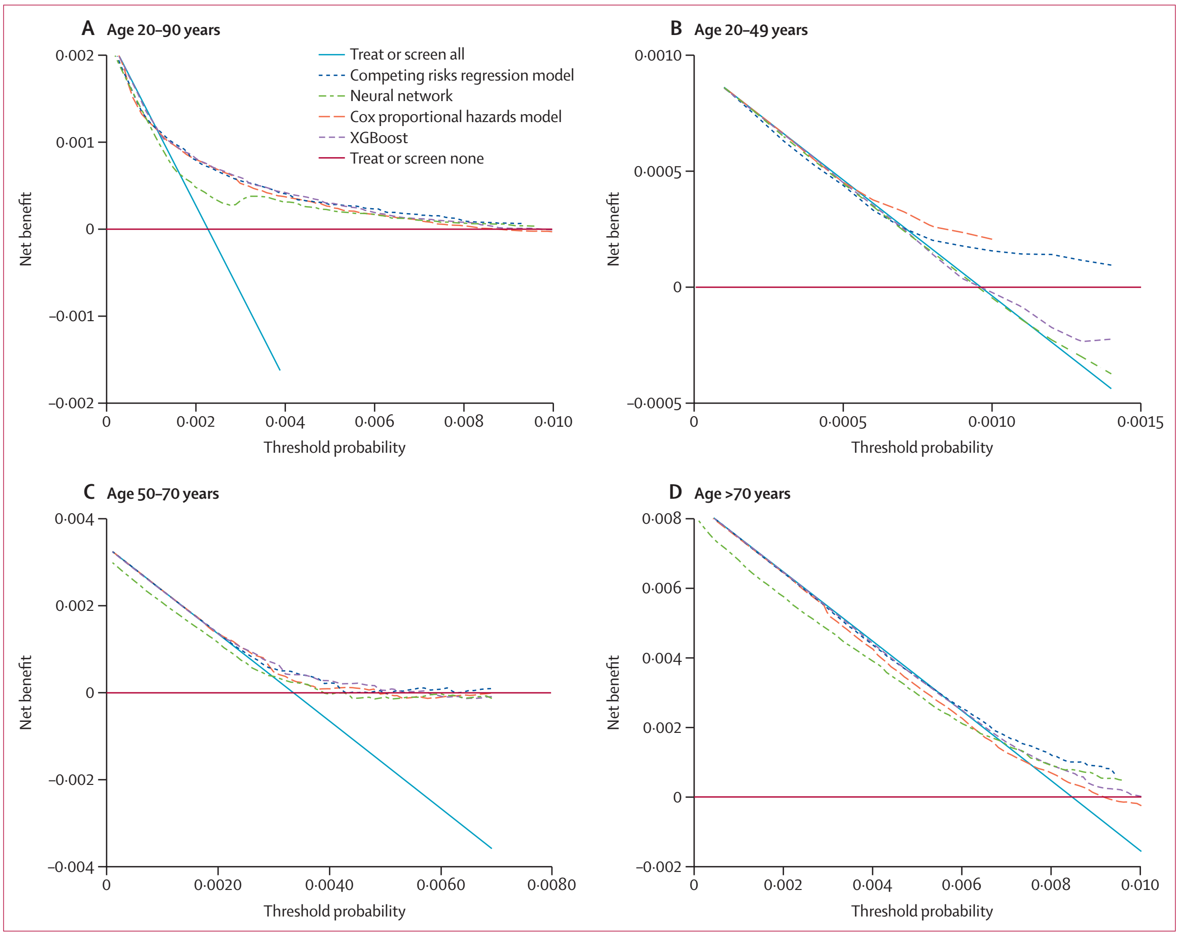

Figure 4: **Clinical utility of each model assessed using net benefit**
Clinical utility was assessed using decision curve analysis, which was done overall (ie, all individuals in period 2), and across age-related subgroups.

with the best net benefit for the post-screening age subgroup. In the screening-age group, all models were associated with a modest difference in net benefit (ie, the results were very similar to the screen allstrategy).

Regarding the models that included ethnicity as a predictor, overall performance metrics were generally similar to the results obtained when ethnicity was not included as a predictor (appendix pp 10, 12). The Cox proportional hazards model that included ethnicity had a Harrell's C index of 0.885 (95% CI 0·842–0·867, 95% prediction interval: 0·821–0·888), compared with 0·854 (0·842–0·865, 0·822–0·885) for the Cox model that did not (appendix p 12).

Discussion

To our knowledge, this study is the largest to develop clinical prediction models in breast cancer yet and is the first to develop models estimating the risks of breast cancer mortality in the general female population. The competing risks model appears the most clinically useful, on the basis of its high discrimination, good calibration overall and across ethnic groups, and association with favourable net benefit across all age subgroups in the decision curve analysis. Potential model uses include identification of women at highest risk of developing life-threatening cancers for chemoprevention, expansion of access to screening, modification of screening strategy, or recruitment into trials for risk-based screening or pan-cancer detection assays.

Breast cancer heterogeneity has important considerations for risk-stratified screening or prevention. First, variation in outcomes across cancer subtypes has motivated the development of subtype-specific models[27] and the assessment of model performance in such subgroups.[28] Some models perform less well than others in predicting more aggressive forms of the disease, such as triple negative breast cancers.[28] Second, the IBIS-I and IBIS-II trials show that chemoprevention agents have differential effects against disease subtypes: overall breast cancer incidence was reduced by more than a third with tamoxifen and anastrozole at 15 years follow-up, but the effects against invasive oestrogen receptor negative cancer incidence, and on breast cancer deaths were not significant.[4,5] The present study had an explicit focus on modelling the risk of breast cancer death, which could avoid some of the complexity of tumour heterogeneity in informing mortality-reducing strategies.

Ethnicity is a protected characteristic and was selected for inclusion as a predictor in both regression models due to its association with breast cancer mortality risk. The inclusion of ethnicity slightly improved the overall performance of the Cox model, and thus its exclusion could be considered a form of omitted variable bias that could underestimate absolute risk in some ethnic groups while overestimating in others. If differences in magnitude and direction of breast cancer risk associations in different ethnic groups reflect underascertainment of breast cancer in some groups (or overascertainment in others), then including ethnicity as a predictor in the model could perpetuate this bias. By contrast, without ascribing any causal interpretation to the model coefficients, all else being equal, non-White females could receive systematically lower risk estimates than White females. When the intended use of these models is to identify individuals for mortality-reducing interventions, this effect could manifest as reduced access to such services for some based on ethnicity. This difference could generate or exacerbate health inequities, despite improving outcomes at the whole-population level. A citizen's jury in 2022 explored views on the use of the QCovid prediction model in public health policy for COVID-19,[29] but its findings are relevant to other scenarios. The consensus was that restricting service access on the basis of ethnicity-influenced risk estimates from models should be unacceptable. This conclusion suggests that any prospective implementation of the models reported here could have reduced validity or acceptability if ethnicity was included. Model coefficients for ethnic groups do not reflect immutable biological characteristics. Rather, they are a proxy reflecting structural or sociocontextual factors alongside any biological effect, if one exists. Cognisant of these factors, we presented the models without ethnicity as the primary results.

Study strengths include the large sample size, the use of linked datasets to ascertain predictors and outcomes, and the evaluation strategy. Limitations include the inability to incorporate genetic risk estimates or mammographic density, due to non-availability in the source datasets. Incorporating these factors has offered incremental value to other models in predicting incident breast cancer diagnosis risks,[30] but their potential effect on the present study's risk trajectory of interest is uncertain. This study sought to develop models that would use routinely collected clinical data only. Exploration of polygenic risk scores for breast cancer death using UK Biobank yielded a model unable to be recommended for use due to low sample size, meaning estimates are probably unstable.[31] Furthermore, although breast density is associated with risk of developing breast cancer, it is not associated with breast cancer mortality.[32] We envisioned developing models that could inform the identification of female individuals at high risk of breast cancer mortality, beyond the current age-based screening eligibility criteria. Those outside of current age eligibility would not have recent mammographic imaging available, and therefore these data could not be included in the models. Other limitations include the rarity of breast cancer deaths in some ethnic subgroups in our sample, and reliance on individual health-care practitioner coding for predictor variables or measurements. There could be misclassification bias for some predictor values—eg, absence of family history was assumed to mean no family history, and positive family history might be more likely to be coded in people with more cases of cancer in their family. Another limitation is that the study only used data from individuals in England, as such, the results might not necessarily generalise to other countries.

In conclusion, this study explored four models to predict 10-year risk of breast cancer mortality in females currently without breast cancer. The competing risks regression model was deemed the most clinically useful. Accurate tools that can identify female individuals at increased risk of developing life-threatening breast cancers could inform efficient targeting of individuals most likely to benefit from chemoprevention, novel screening approaches, or recruitment into trials. This study provides evidence regarding the statistical performance of a new model, but the clinical effect of using this model (and the ways it should be used) to inform risk-based screening or prevention needs further assessment. Implementation of this model requires further evaluation including an external validation, and use outside England would require local validation. Future work should include health economic modelling to ascertain cost-effective intervention strategies informed by the competing risks model, which is underway in this group; qualitative work to understand the acceptability of stratified pathways; and health service considerations on effective but non-disruptive implementation of an algorithm-based approach.

Contributors

AKC conceived the study, developed the protocol and statistical analysis plans, developed, reviewed, and revised clinical and diagnostic code groups, undertook the data cleaning and analyses, interpreted the results, and drafted the manuscript. AKC and JH-C obtained funding. GSC provided expert statistical input into developing analysis plans. SL and DD provided expert clinical input to the project. MB provided expert machine learning input to the project. JH-C provided expert epidemiological input into developing statistical analysis plans, created and reviewed clinical and diagnostic code groups, and provided the linked dataset and analysis environment. GSC, SL, DD, SP, MB, and JH-C contributed to the protocol, provided supervision, interpreted the results, and critically reviewed the manuscript. AKC and JH-C directly accessed and verified the underlying data for this project. Due to database access regulations, only AKC and JH-C had full access to raw study data. All authors were eligible to access the raw study data as they are employees of the University of Oxford, subject to requisite information governance training and data security set-up. All authors accept the final responsibility to submit for publication.

Declaration of interests

SL reports institutional funding for clinical trials for which they are the Chief Investigator or Principal Investigator from Cancer Research UK, Boehringer Ingelheim, Piqur Therapeutics, AstraZeneca, Carrick Therapeutics, Sanofi, Merck, Synthon, Roche, and UK National Institute for Health and Care Research (NIHR); consulting fees from Sanofi, GLG Consulting, Rejuversen, Oxford Biodynamics; payment or honoraria for lectures or presentations from Esai, Prosigna, Roche, Pfizer, Novartis and Sanofi; personal support for attending meetings or for travel from Pfizer, Roche, Synthon and Piqur therapeutics; is a co-founder of MitoRx therapeutics; research funding from Cancer Research UK, Against Breast Cancer, Pathios Therapeutics, and NIHR; and participation on a data safety monitoring board or advisory board for Shionogi. JH-C is an unpaid director of QResearch and is a founder and shareholder of ClinRisk and was its medical director until May 31, 2019; and reports other grant support from the Wellcome Trust and the John Fell Fund. All other authors declare no competing interests.

Data sharing

Data collected for this study cannot be made publicly available to others due to the regulations for access to QResearch data. Regulations regarding access to the QResearch database are detailed at: https://www.qresearch.org/information/information-for-researchers/ and include approval of a proposal by the QResearch Scientific Committee. The protocol has been published previously.[14] Analysis code is available at the following repository: https://github.com/AshDF91/Breast-cancer-mortality-prediction-popn-level-.

Acknowledgments

This project was funded via a Clinical Research Training Fellowship awarded to AKC by Cancer Research UK (C2195/A31310). We also acknowledge personal grant support from Cancer Research UK to DD (C8225/A21133), GSC (C49297/A27294), and JH-C (C5255/A18085). SP receives support as a Senior Investigator for NIHR (NF-SI-0616–10103), and from the NIHR Applied Research Collaboration Oxford and Thames Valley. We acknowledge the contribution of Egton Medical Information Systems (EMIS) practices who contribute to QResearch, EMIS Health, University of Nottingham, and the Chancellor, Masters, and Scholars of the University of Oxford for expertise in establishing, developing, and supporting the QResearch database. The HES datasets and civil registration data are used with permission from NHS Digital who retain the copyright for those data. The HES data used in this analysis are copyright (2021) of the Health and Social Care Information Centre and re-used with the permission of the Health and Social Care Information Centre (and the University of Oxford); all rights reserved. This project involves data derived from patient-level information collected by the NHS, as part of the care and support of patients with cancer. The data are collated, maintained, and quality assured by the National Cancer Registration and Analysis Service, which is part of Public Health England. Access to the data was facilitated by the Public Health England Office for Data Release. The Office for National Statistics, Public Health England, and NHS Digital have no responsibility for the analysis or interpretation of the data.

References

1 Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2021; **71:** 209–49.

2 Clift AK, Dodwell D, Lord S, et al. The current status of risk-stratified breast screening. *Br J Cancer* 2022; **126:** 533–50.

3 Pashayan N, Antoniou AC, Ivanus U, et al. Personalized early detection and prevention of breast cancer: ENVISION consensus statement. *Nat Rev Clin Oncol* 2020; **17:** 687–705.

4 Cuzick J, Sestak I, Cawthorn S, et al. Tamoxifen for prevention of breast cancer: extended long-term follow-up of the IBIS-I breast cancer prevention trial. *Lancet Oncol* 2015; **16:** 67–75.

5 Cuzick J, Sestak I, Forbes JF, et al. Use of anastrozole for breast cancer prevention (IBIS-II): long-term results of a randomised controlled trial. *Lancet* 2020; **395:** 117–22.

6 Esserman LJ, Anton-Culver H, Borowsky A, et al. The WISDOM Study: breaking the deadlock in the breast cancer screening debate. *NPJ Breast Cancer* 2017; **3:** 34.

7 Sherman ME, Ichikawa L, Pfeiffer RM, et al. relationship of predicted risk of developing invasive breast cancer, as assessed with three models, and breast cancer mortality among breast cancer patients. *PLoS One* 2016; **11:** e0160966.

8 Welch HG. Cancer screening—the good, the bad, and the ugly. *JAMA Surg* 2022; **157:** 467–68.

9 Jørgensen KJ, Gøtzsche PC, Kalager M, Zahl PH. Breast cancer screening in Denmark: a cohort study of tumor size and overdiagnosis. *Ann Intern Med* 2017; **166:** 313–23.

10 Harbeck N, Penault-Llorca F, Cortes J, et al. Breast cancer. *Nat Rev Dis Primers* 2019; **5:** 66.

11 Autier P, Boniol M. Mammography screening: a major issue in medicine. *Eur J Cancer* 2018; **90:** 34–62.

12 Hippisley-Cox J, Coupland C. Development and validation of risk prediction algorithms to estimate future risk of common cancers in men and women: prospective cohort study. *BMJ Open* 2015; **5:** e007825.

13 Clift AK, Dodwell D, Lord S, et al. Development and internal-external validation of statistical and machine learning models for breast cancer prognostication: cohort study. *BMJ* 2023; **381:** e073800.

14 Clift AK, Hippisley-Cox J, Dodwell D, et al. Development and validation of clinical prediction models for breast cancer incidence and mortality: a protocol for a dual cohort study. *BMJ Open* 2022; **12:** e050828.

15 Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015; **350:** g7594.

16 QResearch. QCode group library. 2023. https://www.qresearch.org/data/qcode-group-library/ (accessed Feb 1, 2023).

17 Austin PC, van Klaveren D, Vergouwe Y, Nieboer D, Lee DS, Steyerberg EW. Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. *J Clin Epidemiol* 2016; **79:** 76–85.

18 White IR, Royston P. Imputing missing covariate values for the Cox model. *Stat Med* 2009; **28:** 1982–98.

19 Rubin DB. Multiple imputation for nonresponse in surveys. New York, NY: John Wiley & Sons, 1987.

20 Graw F, Gerds TA, Schumacher M. On pseudo-values for regression analysis in competing risks models. *Lifetime Data Anal* 2009; **15:** 241–55.

21 van der Ploeg T, Datema F, Baatenburg de Jong R, Steyerberg EW. Prediction of survival with alternative modeling techniques using pseudo values. *PLoS One* 2014; **9:** e100234.

22 IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Med Res Methodol* 2014; **14:** 25.

23 Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004; **23:** 723–48.

24 Gerds TA, Andersen PK, Kattan MW. Calibration plots for risk prediction models in the presence of competing risks. *Stat Med* 2014; **33:** 3191–203.

25 Cancer Research UK. Breast cancer statistics. 2022. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer (accessed June 29, 2022).

26 Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II—binary and time-to-event outcomes. *Stat Med* 2019; **38:** 1276–96.

27 Mavaddat N, Michailidou K, Dennis J, et al. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am J Hum Genet* 2019; **104:** 21–34.

28 McCarthy AM, Guan Z, Welch M, et al. Performance of breast cancer risk-assessment models in a large mammography cohort. *J Natl Cancer Inst* 2020; **112:** 489–97.

29 Scottish Government. Citizens' jury on QCovid: report on the jury's conclusions and key findings. https://www.gov.scot/publications/citizens-jury-qcovid-report-jurys-conclusions-key-findings/ (accessed June 29, 2022).

30 van Veen EM, Brentnall AR, Byers H, et al. Use of single-nucleotide polymorphisms and mammographic density plus classic risk factors for breast cancer risk prediction. *JAMA Oncol* 2018; **4:** 476–82.

31 Neale Lab. SNP Heritability for Phenotype 40001_C509. 2022. https://nealelab.github.io/UKBB_ldsc/h2_summary_40001_C509.html (accessed Feb 1, 2023).

32 van der Waal D, Verbeek ALM, Broeders MJM. Breast density and breast cancer-specific survival by detection mode. *BMC Cancer* 2018; **18:** 386.