Development and evaluation of clinical prediction models for risk-stratified early detection, prevention and management of breast cancer

Ashley Kieran Clift

Green Templeton College

A thesis submitted for the Doctor of Philosophy in Cancer Science

Hilary Term 2023

Word count: 33,186

For the two loves in my life, without whom this thesis would not have been possible –

My darling Nicole, and the statistical jack-knife

Acknowledgments

Throughout my DPhil, I have been privileged by and am immensely grateful for the support of my supervisors: Julia Hippisley-Cox, Gary Collins, David Dodwell, Simon Lord, Mike Brady, and Stavros Petrou. Since corralling my 'dream team' of supervisors during the initial application process, their robust input, kind challenges and expert guidance have been invaluable to not only my doctorate, but in transforming me into (I hope) a half-decent clinical epidemiologist. I see them not just as supervisors but friends, and I hope that this body of work makes them proud.

I am indebted to the generosity of Cancer Research UK, who funded my DPhil, and thank Kea Hinsley and Michael Youdell who took such good care of our cohort during and after the challenges of a global pandemic.

Lastly, I thank my friends and family for keeping me going and letting me let my hair down (read: wax lyrical over a liberal pour of gin about the wonders of multiple imputation). The escapism of bagging (some) hills in the Scottish rain with my Bothy Boys, or dancing nights with Nix, BewBew, Edmund and Molly kept me sane when I was spending most of my life writing (semi-)functioning code in a darkened room. Uncle duty with my nephew Luca perked me up when I needed it most and provided convenient Christmas-time educational opportunities through modern classics such as 'Neural Networks for Babies'. My simply wonderful sister Amber and Mumzy steered me through many a rut and sustained many a high. I can't appreciate Alan and Bhavana enough for keeping me fed, watered, clothed and laughing. Lastly, I thank Nana and Gaga for getting me into this whole medical thing in the first place – I love and miss them every day.

Three peer-reviewed publications arose from work undertaken in this thesis. In three chapters, some material has been re-used, such as summary tables in **Chapter 1**, descriptions of methodology in **Chapter 2,** and analysis results in **Chapter 4.** This has been stated at the start of relevant chapters, and a list of publications has been supplied below. I wrote the material that was reproduced, am first author on these papers, and the other co-authors (all thesis supervisors) have given permission to use this material in this thesis.

List of publications from this thesis

Clift, AK, et al. The current status of risk-stratified breast screening. Br J Cancer 2022; 126(4):533-550.

Clift, AK. 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

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

Table of contents

Abstract

Accurately estimating individual-level risks of breast cancer incidence and mortality could inform stratified approaches to screening, prevention, and management that help reduce deaths from breast cancer and are more cost-effective. Increasing interest in 'machine learning' techniques and the integration of phenotypic and genetic data present uncertainty around the best approaches to prognostication in these settings.

In a scoping review, evidence deficiencies were highlighted regarding risk-based breast screening, and parallels explored between this and risk-based breast cancer prevention and management. Using the QResearch primary care database and its linked datasets, clinical prediction models were developed using regression and machine learning methods to estimate individual women's 10-year risks of incident breast cancer, 10-year risks of developing and then dying from breast cancer, and 10-year risks of breast cancer mortality after diagnosis. These were comparatively evaluated in an internal-external cross-validation framework to assess performance and transportability. Second, the incremental effects of integrating genome-wide polygenic risk scores to models for the first two outcomes were assessed using the UK Biobank.

This thesis sought to develop and evaluate models in accordance with modern practice, with an emphasis on robust, fair comparisons between different approaches. A key output was the first work to develop models that estimate the risk of breast cancer mortality in women without breast cancer at baseline, which could be important in the context of screening-associated overdiagnosis. Second, models were produced that may provide reliable risk stratification for any woman with breast cancer. Third, integrating genome-wide polygenic risk to 'phenotypic' information may increase the performance of breast cancer incidence and mortality models in women currently eligible for screening. The

clinical impact and cost-effectiveness of strategies structured around these models needs further assessment.

List of Figures

List of Tables

List of Abbreviations

AUC – area under the receiver operating curve

BCSC – Breast Cancer Surveillance Consortium

BI-RADS – Breast Imaging Reporting and Data System

BMI – body mass index (kg/m^2)

CI – confidence interval

DCIS – ductal carcinoma *in situ*

ER – oestrogen receptor

HER2 – human epidermal growth factor 2 receptor

HR – hazard ratio

HRT – hormone replacement therapy

ICER – incremental cost-effectiveness ratio

IECV – internal-external cross-validation

MRI – magnetic resonance imaging

NHS – National Health Service (United Kingdom)

O/E – observed to expected

PI – prediction interval

PR – progesterone receptor

PROBAST – Prediction model risk of bias assessment tool

PRS – polygenic risk score

QALY – quality-adjusted life year

SNP – single nucleotide polymorphism

USPTF – United States Preventive Task Force

Chapter One

Introduction to risk stratified breast screening and thesis motivation

Parts of this chapter have been peer reviewed and published previously elsewhere:

Clift, AK, et al. The current status of risk-stratified breast screening. Br J Cancer 2022; 126(4):533-550.

Summary

This chapter introduces the clinical problem presented by breast cancer and the rationale for considering novel paradigms for screening based on individual women's risks. It then describes the methodology and results from a critical scoping review undertaken to understand the multi-disciplinary evidence landscape relevant to the current status of risk-stratified screening. It then concludes by discussing the four key motivations for the work undertaken in this thesis and outlines the two work packages that it entails.

Introduction

Despite significant improvements in treatment over recent decades, and the availability of population-level early detection programmes in many countries, breast cancer exerts a significant global health burden with over 685,000 deaths per year[1]. In the United Kingdom, breast cancer is the commonest cancer affecting women, with over 55,000 diagnoses and 11,000 deaths annually[2].

Early detection strategies typically comprise screening mammography – this seeks to reduce breast cancer mortality through expedited diagnosis of smaller, asymptomatic/pre-symptomatic lesions. Whilst generally starting in middle age, screening eligibility criteria and practices vary by country (**Table 1.1**). In rare, very high-risk scenarios such as known carriage of a high-penetrance genetic predisposition, earlier screening with magnetic resonance or ultrasound imaging is advocated[3,4].

Meta-analyses of randomised clinical trial data demonstrate reductions in the relative risk of breast cancer mortality due to screening[5], but reduced breast cancer deaths may come at the expense of overdiagnosis (defined as the identification of clinically insignificant tumours which are then treated)[6-9], as well as the consequences of false positive or false negative results[10]. This expense is borne medically and psychologically by women, and financially by healthcare providers. Mirroring the breadth of effect sizes from different systematic reviews of the effect of screening[5,11] which may be influenced by authors' expertise and potential conflicts of interest[12], there is wide variability in the results seen in observational studies or estimated from trials regarding the extent of overdiagnosis, depending on the analytical or modelling approaches used[6,7,9,13-18].

According to the summation of evidence by the UK Independent Panel, breast cancer screening does reduce breast cancer mortality – the pooled relative risk from 11 trials was

0·80 (95% confidence interval, [CI]: 0·73-0·89) – but it has risks[5]. It concluded that for every 10,000 women aged 50 years invited to screening for the next 20 years, 43 breast cancer deaths would be avoided and 681 additional tumours (either invasive, or ductal carcinoma *in situ* [DCIS]) would be diagnosed, but 129 women would be overdiagnosed. This represents 3 overdiagnosed cases per breast cancer death, although this benefit-to harm balance was contested[19]. The UK's breast screening programme was initiated in 1988 following the publication of the Forrest report[20] in 1986, which reviewed evidence from the Swedish 'Two Counties'[21] and the New York Health Insurance Plan[22] trials. It initially invited women aged 50-64 years of age for triennial screens, and the age range was extended up to 70 years in the mid-2000s[23]. Most countries use an age-based population screening strategy that inherently does not account for the fact that the risks of breast cancer incidence and mortality are not equally distributed throughout the female population. In this context, the occurrence of overdiagnosis[5], evidence that age-based service screening programmes may not be cost-effective[24], and the fact that treatments have improved which may narrow the window of opportunity for screening to improve outcomes, the concept of 'risk-stratified screening' has emerged in recent decades[25-27]. Herein, decisions to offer screening, determining screening frequency, modality or starting age would be influenced by individualised risk estimation. Targeted intensification of early detection or prevention measures in higher risk women (e.g. reduced time intervals between screens or changing radiological modality), and reducing (or even stopping) screening in women that stand to gain little benefit could yield more efficacious and cost-effective strategies, should individual level risk estimation mechanisms be accurate enough and the risk-informed interventions be appropriate.

Country	Age group	Screening strategy
United Kingdom (NHS Breast Screening Programme) [28]	Women aged 50-70 years Women aged 71 years and older	Invitation to mammography screening every 3 years Not invited - may self-refer
United States of America (United States Preventive Service Task Force) [29]	Women aged 40 to 49 years Women aged 50 to 74 years Women aged 75 years and older	Individual decision-making recommended Biennial mammography No recommendation: evidence insufficient to assess harms and benefits in this age group
Canada (Canadian Task Force on Preventive Health Care) [30]	Women aged 40 to 49 years Women aged 50 to 69 years Women aged 70 to 74 years	Not recommended; shared decision making if desired Mammography every 2 to 3 years Mammography every 2 to 3 years
Netherlands (National Brest Cancer Screening Programme) [31]	Women aged 50 to 75 years	Invitation to mammography every 2 years
Australia (BreastScreen Australia) [32]	Women aged 40 to 49 years Women aged 50 to 74 years Women aged 74 years and older	Not invited, but may 'opt-in' Invitation to mammography every 2 years Not invited but may 'opt-in'
China (National Health Commission of the People's Republic of China) [33]	Women aged 20 to 39 years Women aged 40 to 69 years Women aged 70 years and older	Monthly breast self-examination, clinical breast examination 1-3 yearly Mammography every 1-2 years with ultrasound for women with dense breasts; monthly breast self-examination and annual clinical breast examination Monthly breast self-examination, annual clinical breast examination

Table 1.1. Summary of selected national screening programme strategies or national body recommendations for screening women that are not at elevated risk of breast cancer (e.g. those without a known familial risk/genetic predisposition, or history of chest wall radiotherapy).

Assessing the feasibility and potential effectiveness of risk-stratified screening requires triangulation of evidence from several fields. This includes but is not limited to: studies developing and evaluating prediction models, epidemiological studies, clinical trials, health economic evaluation, and qualitative analyses. To map this inter-disciplinary evidence landscape, a scoping literature review was undertaken.

Scoping review methodology

A scoping literature review was undertaken using Medline (PubMed) with the following search strategy: ("breast screening" OR "mammography") AND ("risk#adapted" OR "risk#stratified" OR "personalised" OR "personalized" OR "tailored" OR "risk#based"). Papers published in any language prior to 1st November 2020 were considered for inclusion. The reference lists of systematic reviews were reviewed to identify key publications if not identified by the search strategy. The website clinicaltrials.gov was also searched on 1st November 2020 to identify ongoing interventional studies in this area (search terms: "breast cancer", "screening" and "risk"). Reports retrieved were screened for inclusion based on title and abstract and, if relevant, were classified into five groups:

1) Papers reporting risk prediction models
2) Epidemiological analyses of risk-stratified screening or retrospective evaluations
3) Prospective studies and trials of risk-stratified screening
4) Health economic evaluations
5) Qualitative research on feasibility or acceptability.

Findings were synthesised narratively, informed by the narrative synthesis guidelines developed by the Cochrane Collaboration[34]. This search string was re-run periodically throughout the DPhil (every 3 months, last time 1st April 2023) to inform thesis write-up with updated information as relevant.

Risk prediction models to guide personalised screening

Multivariable clinical prediction models provide estimates of the absolute risk of a clinical event of interest by combining the 'effects' of several factors – this can be performed using mathematical modelling[35], statistical regression[36] or 'machine learning' approaches[37].

Risk-stratified screening is typically envisioned to be based around estimation of individual women's risk of developing breast cancer (i.e. being diagnosed). Several tools modelling this endpoint have been published, generally incorporating predictors such as age, pharmacological exposures such as use of hormone replacement therapy, body mass index, family history of breast cancer, and reproductive history. More recently, groups have sought to integrate imaging-related parameters[38], and/or summations of genomic risk[38-40] into 'clinical' risk models.

The 'IBIS' model (Tyrer-Cuzick) can incorporate breast cancer genes 1/2 (*BRCA)* phenotype, and later updates have examined the effects of including a polygenic risk score comprising 313 single nucleotide polymorphisms[38]. Some models are predominantly based on genetic information, e.g. BOADICEA[41] and BRCAPRO[42,43]. Highly penetrant mutations such as those in *BRCA1* or *BRCA2* are relatively rare[44,45], and it has been

estimated that only 25-50% of the familial risk of breast cancer is accounted for by known genetic variants[44,46-49], and 16% of the risk of non-familial breast cancer is accounted for by single nucleotide polymorphisms (SNP) panels[45]. Imaging-derived parameters such as textural features or breast density have been of interest[50] – the latter can be determined with visual assessment scales or automated approaches[51,52], most commonly with the four-category 'Breast Imaging Reporting and Data System' classification (BI-RADS)[53,54].

Although the implementation of risk-stratified screening strategies to a population should arguably follow prospective outcome and economic evaluations, an accurate risk estimation mechanism is the *sine qua non*. Any clinical prediction model should be transparently reported to permit review and promote confidence in end-users, and robustly evaluated in terms of various performance metrics[55]. Strong internal validation (i.e. in the derivation data) is recommended, and external validation using data sources independent from the data used to develop the model can provide useful information[56]. It is important to consider model discrimination (how well the model distinguishes between events and non-events), calibration (accuracy of risk predictions, alignment between predicted and observed probabilities), and clinical utility (such as 'net benefit') [57,58]. Net benefit approaches extend the evaluation of a model to consider the trade-offs between positive outcomes (e.g. detecting a cancer) versus negative outcomes (e.g. missed diagnoses) if a model was used to inform clinical decision making[59]. These trade-offs are typically estimated across a range of 'threshold probabilities' (model-derived probabilities at which an intervention would be triggered), and then plotted as a 'decision curve'. This 'decision curve analysis' implicitly takes into account discrimination and calibration – for example, a poorly calibrated model will produce unreliable estimates and trigger decisions in a way that distorts the trade-off between benefits and harms. The

strategy that has the highest net benefit at a given threshold is deemed the most clinically valuable[59].

Depending on the modelling strategy, other metrics such as Royston & Sauerbrei's R^2 may be calculable, which estimates the variation in time-to-event captured by a Cox proportional hazards model[60]. Moreover, it is increasingly apparent that the summation of performance with 'overall' estimates of these metrics can be insufficient – models may have different performance characteristics in societally or clinically relevant groups defined by ethnicity, age-groups, or disease sub-type[61,62].

Another type of model performance metric that can be considered is the net reclassification index (NRI)[188]. This typically aims to quantify the incremental improvement of adding a new predictor or marker to a model's performance, based on its ability to correctly (re)classify individuals into risk groups. It can also be used to compare two different models. However, use of the NRI has attracted significant criticism[189,190] – it does not account for the relative severity of different types of error, shows instability when comparing models that are poorly calibrated, and is dependent on the choice of risk groups. For the latter, risk grouping may be an arbitrary dichotomisation, could be 'gamed' by researchers, or there may be no consensus on what an optimal risk group structure is (or if one even exists). Due to these limitations, NRI is not considered further in this thesis.

Table 1.2 summarises important methodological aspects and results from key published risk prediction model studies.

Study first author and year	Study design and setting	Outcome modelled	Study size (n) and cases	Risk model predictors / data type combinations	Validation strategy	Discrimination metrics reported (95% confidence interval)	Calibration metrics reported (95% confidence interval)	Examination of performance heterogeneity
'Tyrer-Cuzick' or 'IBIS' model and updates or variations								
Tyrer, et al., 2004 [35]	Construction of model from first principles/ informed by published data	Diagnosis of breast cancer	N/A	Age, BRCA genotype, family history of breast cancer, (including relationship and ages), menarche, age at first birth, menopausal status, atypical hyperplasia, lobular carcinoma in situ, height, BMI	None	None	None	None
Brentnall, et al., 2018 [38]	Cohort study (Kaiser Permanente Washington registry), US, women aged 40-73 years	Diagnosis of breast cancer within 19 years	132,139 (2,699 breast cancer cases)	Tyrer-Cuzick model Tyrer-Cuzick model + breast density category	Evaluated pre-designed risk calculation mechanism in the whole cohort	Separation of cumulative risk curves for pre-determined risk groups Upper tenth, lower tenth and 'middle' 8 tenths compared with hazard ratios	O/E: 1.02 (0.98-1.06) O/E: 0.98 (0.94-1.02)	O/E of models by 3 age groups (<50years, 50-59years, 60+ years)
van Veen, et al., 2018 [40]	Cohort study (PROCAS), England,	Diagnosis of breast cancer within 10 years	9,363	Tyrer-Cuzick model	Evaluated pre-designed risk	AUC 0.58 (0.52-0.62)		Discrimination and calibration in terms of tumour ER

	women aged 46-73 years		(466 cases, 196 incident)	Tyrer-Cuzick + mammographic density Tyrer-Cuzick + mammographic density + polygenic risk score (18 SNPs)	calculation mechanism in the whole cohort	AUC 0.64 (0.60-0.68) AUC 0.67 (0.62-0.71)	O/E: 1.50 (1.33-1.70) O/E: 0.98 (0.69-1.28)	status, and by invasive/DCIS tumour type
'Gail' or 'BCRAT' model and updates or variations								
Gail, et al., 1989 [63]	Case-control study, US, White women aged 50 years and over	Projected breast cancer probabilities within 10, 20 and 30 years of follow-up	5,998 (2,852 cases)	Age at menarche, age at first live birth, number of previous breast biopsies, number of first-degree relatives with breast cancer	None	None	None	None
Gail, et al., 2007 [64]	Case-control study ('CARE'), US, African-American women aged 35-64 years	Projected breast cancer risk within 10, 20 and 30 years of follow-up	3,254 (1,607 breast cancer cases)	As for BCRAT, re-fitted in data from African-American women	External validation Women's Health Initiative cohort	C-index 0.555 (0.535-0.575)	O/E: 1.08 (0.97-1.20)	O/E ratio by age groups, ae at menarche, number of biopsies, and number of first-degree relatives C-index by age groups
Tice, et al., 2005 [65]	Cohort study, US, women aged 35 years and over	Diagnosis of invasive breast cancer (no explicit horizon, median follow-up 5.1 years)	81,777 (955 breast cancer cases)	BCRAT BCRAT + BI-RADS breast density category	Apparent validation in study dataset	C-index 0.67 (0.65-0.68) C-index 0.68 (0.66-0.70)	None	None

Zhang, et al., 2018 [66]	Nested case-control study, US, women aged 34-70 years	Diagnosis of invasive breast cancer within 5 years	11,880 (4,006 cases)	BCRAT risk score + polygenic risk score (67 SNPs) + mammographic density + estrone sulfate + testosterone + prolactin	Internal: cross-validation	All women: AUC 0.65 (0.64 to 0.66) Described 5-year risk per predicted risk percentile	None	Age-group-specific changes in AUC with sequential addition of variables to model
BCSC model (Breast Cancer Surveillance Consortium)								
Tice, et al., 2015 [67]	Cohort study, US, women undergoing mammography aged 35 years and over	Diagnosis of invasive breast cancer within 5 years, and 10 years	1,135,977 (17,908 breast cancer cases)	Age, race/ethnicity, family history of breast cancer, breast biopsy history, benign breast disease, BI-RADS breast density	Update of earlier model developed using sample of cohort[68], five-fold cross-validation	AUC 0.665	E/O ratio 1.04 (1.03 to 1.06)	E/O across 5year age groups, ethnic groups, levels of predictor parameters
Vachon, et al., 2015 [69]	Nested case-control study, US women undergoing screening; 2 further case-control studies (US and Germany)	Diagnosis of invasive breast cancer within 5 years	1,622 (set 1) 1,529 (set 2) 879 (set 3) (Total 1,634 breast cancer cases)	BCSC + PRS (76 loci) + BI-RADS breast density	Evaluation of models formed using datasets 2&3 using dataset 1	AUC 0.69 (0.67 to 0.71)	Hosmer-Lemeshow test of calibration	None
Rosner-Colditz model and updates or variations								
Rosner & Colditz, 1996 [70]	Cohort study, US, Caucasian	Diagnosis of breast cancer	89,132	Age, age at menarche, age at menopause, age at	None	None	None	None

	women aged 30-64 years	(no specific horizon)	(2,249 breast cancer cases)	first birth, age at subsequent births				
Rosner, et al., 2008 [71]	Cohort study, US, Caucasian women aged 30-64 years	Diagnosis of breast cancer (no specific horizon)	66,145 (1,559 breast cancer cases)	Age, age at menarche, age at menopause, age at first birth, age at subsequent births, benign breast disease history, HRT use, first degree family history of breast cancer, weight, BMI, alcohol intake, oestradiol levels	None (Assessed fit of log-incidence model)	AUC 0.645	None	None
Zhang, et al., 2018 [66]	Nested case-control study, US, women aged 34-70 years	Diagnosis of invasive breast cancer within 5 years	11,880 (4,006 cases)	BCRAT risk score + polygenic risk score (67 SNPs) + mammographic density + estrone sulfate + testosterone + prolactin	Internal: cross-validation	All women: AUC 0.678 (0.666 to 0.690) Described 5-year risk per predicted risk percentile	None	Age-group-specific changes in AUC with sequential addition of variables to model
BOADICEA V6 (multifactorial model)								
Yang, et al. 2022[187]	Prospective cohort study in Sweden (KARMA), women aged 25-74 years	Diagnosis of breast cancer within 5 years	Full cohort: 66,415 (816 breast cancer cases) Sub-cohort with PRS data: 15,502	Age, year of birth (to identify birth cohort), ethnicity, Ashkenazi Jewish heritage, height, BMI, identical twin status, daily alcohol intake, age at menarche, number of live births, age at first birth, menopausal	External validation in prospective cohort	(Metrics are reported for the specific model with maximal number of predictors) Full cohort (without	(Metrics are reported for the specific model with maximal number of predictors) Full cohort (without PRS/pathog	Reported discrimination and calibration statistics with sequential addition of model components up to maximally complex multifactorial model

	(767 breast cancer cases) Sub-cohort with pathogenic variant status information 5,693 (280 breast cancer cases)	status, age at menopause, oral contraceptive pill use, HRT use, history of ovarian cancer, mammographic density, *BRCA1/2* mutation status, *ATM/CHEK2/PALB2/RAD51C/RAD51D/BARD1* mutation status, PRS with 313 SNPs, 'residual polygenic risk' component, family history of breast cancer, contralateral breast cancer, ovarian cancer, pancreatic cancer, prostate cancer, age at cancer diagnosis (in relations)	PRS/pathogenic variant status): Harrell's C 0.64 (0.62 to 0.66) Sub-cohort with PRS information: Harrell's C 0.69 (0.67 to 0.71) Sub-cohort with PRS and pathogenic variant information: Harrell's C 0.71 (0.68 to 0.74)	enic variant status): Calibration slope 0.98 (0.96 to 1.00) E/O ratio 0.92 (0.86 to 0.99) Sub-cohort with PRS information: Calibration slope 0.98 (0.96 to 1.00) E/O ratio 0.92 (0.85 to 0.99) Sub-cohort with PRS and pathogenic variant information: Calibration slope 0.97 (0.95 to 0.99) E/O ratio 0.88 (0.75 to 1.04) Reported calibration plots	Also reported model performance by menopausal status (i.e. pre- and post-menopausal)

							grouped by tenths of predicted risks	
Other models								
Hippisley-Cox & Coupland, 2015 [36]	Cohort study in primary care (England), women aged 25-84 years	Diagnosis of breast cancer within 10 years	3,318,258 (41,315 breast cancer cases)	Age, BMI, deprivation score, ethnicity, alcohol intake, family history of breast cancer, benign breast disease, oral contraceptive use, oestrogen-containing HRT use, manic depression/schizophrenia, previous blood cancer, previous lung cancer, previous ovarian cancer	Split-sample validation	C-index 0.761 (0.758 to 0.765) D statistic 1.088 (1.058 to 1.119)	Calibration plots (by tenth of predicted risk)	Performance by age groups
Pal Choudhury, et al., (iCARE models), 2020[72]	Cohort study in UK, women aged 35-74 years	Diagnosis of invasive breast cancer within 5 years	UK-based validation: 64,874 (863 breast cancer cases)	*iCARE-BPC3*: age at menarche, age at menopause, parity, age at first birth, height, alcohol intake, first-degree relative with breast cancer, ever/never smoker, BMI, HRT use (type), current HRT use, interaction between BMI and HRT type	External validation of models	*UK:* AUC over 50s: 0.602 (0.580 to 0.624)	Over 50s: E/O: 1.00 (0.93-1.09) CS: 0.960 (0.680-1.239) CI: 0.001 (-0.004-0.005)	Discrimination and calibration metrics by age groups (<50yr, 50+yrs)

			US-based validation for iCARE-Lit: 47,279 (1008 breast cancer cases)	*iCARE-Lit*: age at menarche, age at menopause, parity, age at first birth, OCP use, BMI, height, alcohol intake, benign breast disease history, first degree family history of breast cancer, interaction between BMI and HRT use		AUC under 50s: 0.654 (0.621 to 0.687) AUC over 50s: 0.622 (0.600 to 0.645)	Over 50s: E/O 1.13 (1.04-1.22) CS: 0.811 (0.668-0.954) CI: 0.001 (-0.001-0.004)	
Ming, et al., 2020 [37]	Single centre, women aged 20-80 years, Switzerland	Lifetime risk of breast cancer (Grouped into <17%, 17-29%, 30+ %)	112,587 (4,911 breast cancer cases)	ML classification models: Markov Chain Monte Carlo generalised linear mixed model, adaptive boosting and random forest Same inputs as BOADICEA	Internal: repeated cross-validation	ML models: AUC 0.843 to 0.889 BOADICEA: AUC 0.639	Not presented	None
Park, et al., (KoBCRAT), 2013 [73]	Case-control study, Korea No age limits	5-year and lifetime risk of breast cancer	9,248 (4,601 breast cancer cases)	Same predictors as Gail model, re-developed in Korean population	External validation in further Korean cohort studies	AUC 0.61 (0.49 to 0.72) in one cohort, AUC 0.89 (0.85 to 0.93 in the other)	E/O: 0.97 (0.67-1.40) 1st cohort, 0.96 (0.70-1.37) in 2nd cohort	AUCs in women aged <50yrs and 50+yrs
Abdolell, et al., 2020 [74]	Nested case-control study from population-based screening program in Canada,	Risk of breast cancer diagnosis at mammography	7,770 (1,882 breast cancer cases)	Percentage mammographic density, breast volume, age, cope biopsy history, first-degree family history, number of	Apparent performance in study dataset (no validation)	Imaging features model: AUC 0.597 Imaging features + biopsy history: AUC 0.660	Not presented	None

	women aged 40-75			births, menopausal status, HRT use		Full model: AUC 0.665		
Wang, et al., 2014 [75]	Systematic review to pool odds ratios from observational studies to produce risk score. Evaluated in Chinese women	Risk of breast cancer within 5 years	62,875 (15 cases of breast cancer)	Age at menarche, age at first birth, benign breast disease history, family history of breast cancer, history of breastfeeding, number of terminations of pregnancy	Evaluated pre-designed risk calculation mechanism in the whole cohort	AUC 0.64 (0.50-0.78)	None	None
Ueda, et al., 2003 [76]	Case-control study, Japanese women at single institution, aged	Risk of breast cancer within 10 years, 20 years, and until age 84 years	376 breast cancer patients (no information on number of controls)	Age at menarche, age at first delivery, family history of breast cancer, BMI (if post-menopausal)	None	None	None	None
Eriksson, et al., 2017 [77]	Nested case-control study, Swedish women aged 31-79 years undergoing screening	Risk of breast cancer within 2 years	2,165 (433 cancer cases)	*(Different model weights for pre- and post-menopausal women):* BMI, HRT use, family history of breast cancer, breast density, absolute difference in density between breasts, absolute difference in microcalcification between breasts, interaction term	Cross-validation (number of folds not specified)	AUC 0.71 (0.69-0.73)	None	None

				for % breast density and masses				
Barlow, et al., 2006 [78]	Cohort study, US women aged 35-84 years who underwent a mammogram in the preceding 5 years	Risk of invasive breast cancer or DCIS following a screening mammogram within 1 year	1,007,600 (11,638 breast cancer cases)	*Pre-menopausal women:* Age, breast density (BI-RADS category), family history of breast cancer, and a prior breast procedure *Postmenopausal women:* Age, breast density, race/ethnicity, family history of breast cancer, prior breast procedure, BMI, natural menopause, HRT, prior false-positive mammogram	Split sample validation	Pre-menopausal women: AUC 0.629 (0.603-0.656) Post-menopausal women: AUC 0.626 (0.615-0.637)	Hosmer-Lemeshow goodness of fit	None (separate models fitted for pre- and post-menopausal women)

Table 1.2. Details regarding study data, modelling strategy and performance metrics of notable published risk prediction models for breast cancer or their 'updates' identified during the scoping review. N/A = not applicable, O/E ratio = observed to expected ratio, E/O ratio = expected to observed ratio, BI-RADS = Breast Imaging – Reporting and Data system classification, AUC = area under the receiver operating curve, BMI = body mass index, HRT = hormone replacement therapy, ML = machine learning, ER+ = oestrogen receptor positive, ER- = oestrogen receptor negative, PR+ = progesterone receptor positive, PR- = progesterone receptor negative, SNP = single nucleotide polymorphisms, CS = calibration slope, CI: calibration intercept (i.e. calibration-in-the-large). In terms of discrimination metrics, the c-statistic or c-index is reported: as the AUC and c-statistic are identical for binary outcomes, whereas c-index is specifically for survival/time-to-event data, I have named the metrics using the appropriate nomenclature dictated by statistical approach used, regardless of the name used in the original publication. It is important to note that the AUC/c-indices for each model development

paper are not directly comparable, as each study population varies by age, region and setting. The above information is intended as a narrative overview of extant models as they are reported in their respective study populations and does not constitute a formal comparison of model performance.

At initial search, the systematic review of Louro, et al.[79] was the most recent to appraise the evidence for clinical prediction models for incident breast cancer risk. Using the ISPOR-AMCP-NPC framework to assess studies[80], Louro and colleagues assessed models including the BCRAT, BCSC, IBIS, and concluded it could not recommend any model for the purposes of risk-stratified screening due to methodological limitations or uncertainties of performance[79]. Since then, a systematic review by Zheng, et al.[81] used PROBAST[82] to assess 47 breast cancer incidence models from 40 studies published until December 2021. Most models were developed using data solely from Caucasian women, only nine included internal validation, and all were at high risk of bias[81]. However, neither study included the QCancer-10year (Breast) model, which had the largest sample size yet used to develop and validate a prediction model for incident breast cancer[36]. The emergent pattern is that integration of multimodal data, e.g. phenotypic and genetic, yield incremental gains in model performance[38,40,83-85,191,192], although these are typically small (e.g. increase in AUC by 0.03 to 0.14[86]). In some studies, the effects of integrating multimodal data are measured only using discrimination indices with unclear effects on calibration or net benefit.

Comparing model development and validation studies can be non-trivial – the samples used therein may differ in terms of age profiles or other forms of heterogeneity, the prediction horizons used, and exclusions made. For example, the discrimination of a model assessed in a cohort with a broad age range is not directly comparable to the performance of a model tested in a separate cohort with a narrower one, particularly as a significant proportion of risk is driven by increasing age. However, the comprehensiveness of evaluation is more comparable. The observed-to-expected ratio (O/E) or its converse (E/O) is widely used to assess calibration by comparing the total number of observed events to the number predicted by a model, but is insufficient as a

standalone metric as over-prediction in some sub-groups can be compensated by under-prediction in others. More comprehensive analysis of model calibration is possible with calibration plots which display (mis)alignment across the spectrum of risks[87]. Other metrics that have been used include the relative risks of pre-selected highest risk and lower risk groups, such as the top 10%, middle 80%, and lowest 10% of predicted risks[38]. This alone is an incomplete assessment of discrimination – smaller groups at the extremes of predicted risk distribution are expected to be divergent from the bulk of the study sample. Model performance should be interpreted in terms of the metrics used and the sources of the study data.

A key external validation study of four models in a cohort of 15,732 women aged 20-70 years from Australia, Canada and the US (519 cases of breast cancer) demonstrated c-statistics of 0.70 for BOADICEA (95% CI: 0.68-0.72), 0.71 for IBIS (95% CI: 0.69-0.73), 0.68 for BRCAPRO (95% CI: 0.65-0.70) and 0.60 for BRCAT (0.58-0.62)[88]. Assessment of calibration was limited to the O/E ratio: BOADICEA 1.05 (95% CI: 0.97-1.14), IBIS 1.03 (95% CI: 0.96-1.12), BRCAPRO 0.68 (95% CI: 0.65-0.70), and BCRAT 0.79 (95% CI: 0.73-0.85)[88].

There has been increasing interest in 'machine learning' prediction modelling for healthcare purposes[89-91]. Whilst perceived to be more flexible than some statistical approaches (e.g. better at capturing non-linearities without explicit programming, or more complex/higher-order interactions), less concerned with assumptions on data-generating mechanisms, and capable of handling some forms of data that regression models cannot (e.g. imaging), no single machine learning approach has been shown to be inherently superior to any other [92-95]. The reliability of comparisons between 'traditional' and machine learning approaches using time-to-event, censored data can be problematic – a recent scoping review of 10 simulation studies published between 2000 and 2020 reported

that comparison methodology was poorly transparent, results were often biased towards the 'novel' comparator, and/or did not permit interactions or non-linearities in the Cox models but did so in the flexible methods[96]. Another systematic review of 152 machine learning-based prediction model studies raised concerns regarding the prevalence of poor reporting practices, misleading interpretation ('spin') and extrapolation, and inappropriate recommendations – for example, 77.6% of studies had misleading reporting practices in their abstracts, and only 8.6% cited the relevant reporting guideline[97].

Datasets used for any predictive modelling should capture clinical reality, i.e. reflect the target population, and the architecture of any machine learning algorithm should be reported, given their structural flexibility[98]. As with statistical models, the clarity of reporting can be problematic[99,100 96]. One study in **Table 1.2** compared the performance of the BOADICEA model with a Markov Chain Monte Carlo generalised linear mixed model, an adaptive boosting model and a random forest model developed using data from a single oncogenetic institution in Switzerland that focusses on counselling and testing for hereditary cancer syndromes[37]. Whilst the study concluded that the machine learning models outperformed BOADICEA, no robust evaluation of model calibration was performed, and the comparison was in effect an external validation of BOADICEA versus an internal validation of the new model using the data they were derived from. This used cross-validation with a low number of repeats (n=20), which presumably was used for hyperparameter tuning as well as performance evaluation (not elaborated in paper), an option that is optimistically biased[101]. Further work in this area of comparing different model building strategies for predicting risk should focus on more robust, meaningful comparisons.

Overall, a range of clinical prediction models have been developed that could be used to guide risk-stratified screening, some of which are undergoing evaluation in trials of

personalised screening (see later in this Chapter). There is uncertainty regarding the ability of current models to stratify the asymptomatic female population well enough for risk-based screening, even if multimodal data sources are used. No threshold for a single performance metric will render a model suitable for risk-based screening, mandating a comprehensive approach to model validation. A model with an AUC or C-index of 0.5 cannot be clinically useful, but a high AUC (however defined) does not guarantee clinical utility. Poorly calibrated models may cause harm, and those with unstable performance across sub-groups may raise concerns regarding 'algorithmic fairness/bias'[102,103]. Weak or non-existent calibration assessment, non-examination of performance heterogeneity, or the lack of consideration of geographical and temporal transportability are notable limitations in current models. The QCancer (Breast) [36], IBIS[38] and iCARE[72] models are some examples wherein exploration of performance heterogeneity is performed according to age groups or other clinically relevant sub-populations.

Epidemiological analyses and retrospective evaluations of risk-stratified screening

Assessing the real-world effects of implementing risk-stratified screening strategies may necessitate prospective trial evaluation. Whilst these have been initiated (see later section in this Chapter), several explorations of the possible benefits and harms using epidemiological approaches have been performed, such as in Taiwan or Sweden, alongside retrospective 'simulations' of the effects of implementing different screen practices in screened cohorts. **Table 1.3** summarises key outputs from identified studies.

A large Taiwanese study covering over 1.4 million individuals exploited the natural experiment of the concurrent availability of three screening approaches in the country's population: annual clinical breast examination (baseline, women aged 35+ years), risk-stratified biennial mammographic screening, or universal mammography (both for women aged 50-69 years) [104]. This co-existence of three strategies was predicated by a low breast cancer incidence in 2002-04 (which has increased since) [105], and contemporary concerns regarding the health system's capacity for whole population screening. Risk stratification was informed by a score based on reproductive/menstrual history and family history data during attendances for clinical breast examination between 1999 and 2001 – the median of the risk score distribution was used as the cut-off for biennial mammography eligibility. Using propensity score methodology to account for baseline differences in age at menarche, parity, breastfeeding history, and body mass index (BMI), Cox proportional hazards modelling was used, with screening modality as a time-dependent covariate, to estimate hazard ratios (HR) for various endpoints across the three groups. Compared to clinical breast examination, universal biennial mammography was associated with a downwards stage migration of detected cancers with a 30% reduction in stage II+ cancers (HR 0.70, 95% CI: 0.66 to 0.74), a 41% reduction in breast cancer mortality (95% CI: 27 to 52%) after adjustment for propensity score and year of birth and a 13% (95% CI: 8 to 18%) overdiagnosis rate[104]. The overdiagnosis with risk-based screening compared with clinical examination was negligible (hazard ratio for diagnosis 0.97, 95% CI: 0.92 to 1.03), there was an 8% reduction in stage II+ breast cancers (HR 0.92, 95% CI: 0.86 to 0.99), and a non-significant reduction in breast cancer mortality of 14% (HR 0.86, 95% CI: 0.73 to 1.03)[104]. However, the risk stratification model was unclear – data were not provided on the modelling methodology used, the exact risk score covariates, the risk score distribution in the population seeking to opt into risk-based

screening (or across the three groups), nor was there any performance evaluation of this model to assess if it was suitable to guide clinical decision making in the first place. Further detail is needed to make meaningful inference on the performance of risk-based screening versus 'standard' screening in this analysis. Additionally, given the relatively low, albeit increasing, breast cancer incidence rate in this population, risk stratification beyond age and sex may have different proportional benefits in Taiwan in comparison to other nations.

Rather than analysing the effects of altering screening intensity or avoiding screening in low-risk women, a large Swedish cohort with linkage to several national databases comprising over 5,000,000 women was used to assess whether earlier screening starting ages could be appropriate for some women[106,107]. By using 10-year cumulative risk estimations, the risk level of the 'average' 50-year-old woman that would be offered screening was calculated as a benchmark. The ages at which other women would attain the same 10-year risk was compared, based on patterns of family history in one study[107], or personal reproductive history factors[106] (parity, and age at first birth) in another. Both studies found that either approach could identify women that, despite not being eligible to start age-based screening, had the same 10-year risk estimate as 50-year old women that would be invited to screen, or indeed may only attain that same threshold of risk after age 50. For example, women who had their first baby aged under 25 met the benchmark aged 51, whereas women that had four births by age 25 met this at 59 years of age. Furthermore, women with one first-degree relative diagnosed with breast cancer before the age of 40 met the average risk of women starting age-based screening at age 36. Therefore, despite debate around the benefits of universally expanding screening to younger age groups such as the lack of long-term effect seen in Age UK trial[108], selected women with selected prognostic factors may be suitable for earlier or delayed

commencement of early detection strategies. The optimal way to assess risk would however, require further elucidation as a reliance on only two (albeit useful) factors may inadequately capture risk.

In the radiological literature, some commentators have voiced criticism of the potential harms offered by risk-stratified screening, typically fuelled by retrospective studies applying prognostic factor-based decision rules to cohorts of women that partook in service screening[109-111]. For example, Lee and colleagues examined recall rates, cancer detection rates and positive predictive values for biopsy recommendation and 'fact of' biopsy across age groups, when accounting for breast cancer family history, personal breast cancer history, and having dense breasts in cohort of over 2.6 million women[112]. The recall and cancer detection rates in 30-39 year old women were the same with the selected factors undergoing incidence screening as the 40-49 year old 'average risk' women undergoing serial screening; they concluded that such higher risk women may benefit from an earlier screening starting age. Other institutional studies have expressed concern that risk-stratified approaches have the potential to miss 75.6% to 88% of cancers if screening was purely based on family history, 56% to 86% of cancers if density was the sole determinant, or 43.5% to 76% if access to screening was dictated by positive family history and breast density[109,111,113,114]. Such approaches are not powered to assess the effects of varying screening approaches on stage at detection nor breast cancer mortality, but most crucially, they rely on determinations of relative risk far more simplistic than those offered by extant clinical prediction models, such as those being used in prospective trials. Therefore, these studies, rather than undermining the case for considering risk-stratified screening, demonstrate the need for nuanced risk estimation. Overall, currently available epidemiological evaluations or retrospective estimations of risk-based screening are insufficient.

Study	Country and setting	Description of modelling processes	Key results
Retrospective evaluations using clinical data, or epidemiological modelling			
van den Broek, et al., 2020 [115]	US Women aged 30-50 years	Breast cancer simulation models: Average risk women, screened according to USPSTF guideline	*Per 1,000 women screened, during lifetime:* 118 life-years gained, 7 deaths averted, 15 overdiagnoses, 920 false positives
		Family history strategy	125 life years gained, 6.9 deaths averted, 14.9 overdiagnoses, 1000 false positives
		Polygenic breast cancer risk model (313 SNPs)	141 life years gained, 7.4 breast cancer deaths averted, 16.0 overdiagnoses, 1156 false positives
		Family history + polygenic risk	154 life years gained, 7.9 deaths averted, 16.6 overdiagnoses, 1169 false positives
Mukama, et al., 2020 [106]	Sweden, n=5,099,172	10-year cumulative risk curves for breast cancer Analysed risk levels of women with parity and first birth age: risk-adapted starting age of screening based on reproductive profiles	Women with first birth at age <25 years and one child attained same level of risk as average 50-year-old female at age 51, those with parity of 4 or more met this threshold at 59
Mukama, et al., 2019 [107]	Sweden, n=5,099,172	Modelled age at which women with specific permutations of family history variables attained the risk level of the average 50-year women (age at which screening starts)	If screening would be advised to start at 50 years, women with 1 first-degree relative diagnosed with breast cancer aged <40, met the benchmark level of risk at 36 years of age
Lee, et al., 2020 [112]	US Screening mammograms from 2,647,315 women	Separated women into risk groups based on 5-year age bracket, family history of breast cancer, personal history of breast cancer, and dense breasts	Women aged 30-34 years had similar cancer detection rates and recall rates as those aged 40-49 years, suggesting earlier screening in women at higher risk may be appropriate
Burnside, et al., 2019 [113]	US Screening mammograms from 10,280 women	Cross-sectional study comparing two scenarios: standard age-based screening versus risk-based, defined as having 5-year risk greater than average 50-year old	Age-based screening diagnosed more cancers than risk-based (68% vs. 26%), more false positives (50.3% vs. 12.1%)

Prospective epidemiological studies			
Yen, et al., 2016 [104]	Taiwan, population-based cohort study, n=1,429,890	Cohort study of three screening strategies, adjusting for propensity score for participation: clinical breast examination, risk-based mammography and universal mammography (aged 50-69)	No overdiagnosis compared to clinical examination for risk-based screening, versus 13% overdiagnosis with universal screening 41% reduction in breast cancer mortality with universal screening (adjusted for year of birth and propensity score), non-significant reduction with risk-stratified screening

Table 1.3. Comparison of studies evaluating risk-stratified screening using simulations on retrospective data, or epidemiological studies. USPSTF = United States Preventive Service Task Force, SNP = single nucleotide polymorphism.

Prospective cohort studies

Three notable cohort studies are currently exploring the feasibility and acceptability of personalised risk assessment in the general population: specifically, the Personalised RISk-based MAmma Screening study (PRISMA), the Karolinska Mammography Project for Risk Prediction of Breast Cancer (KARMA), and the Predicting the Risk of Cancer at Screening study (PROCAS).

PRISMA is a Dutch collaboration between institutions including Radboud University Medical Centre Nijmegen and the North, East, West and South Screening Programmes. In 2014, PRISMA started recruiting asymptomatic women aged 50-75 in the general population eligible for the national screening programme for data collection via questionnaires, blood and saliva samples, and mammograms for assessing breast density. It has a target of 90,000 women with regards to risk factor questionnaire data collection and imaging, and 27,000 for blood samples. It aims to not only develop risk prediction models as a fulcrum for investigating risk-based screening strategies but undertake robust assessment of the acceptability of risk-based screening from ethical, psychological, legal and logistical perspectives. Notable outputs from PRISMA thus far include results from multi-cohort qualitative research incorporating individuals from KARMA and PROCAS, which identified a preference for risk-tailored assessment result communication (e.g. letters for below average and average risk, face-to-face appointments for higher risk), the need for standardised risk assessments within national policies, and detailed information needs for women in different European countries[116-118].

The KARMA prospective screening cohort is developing an extensive resource of banked biological, mammogram and lifestyle/clinical factor information from over 70,000 women[119]. Its aims include identification of novel circulating risk markers, genetic risk

factors and imaging protocols, assessment of high-throughput breast density measurement, trials of pharmacological prevention therapies such as lower-dose anti-oestrogen therapy, risk communication, as well as the development of new risk prediction models with evaluation of how these could be implemented within screening routines. Notable outputs include the CAD2Y model which integrates mammographic features such as computer-detected micro-calcifications with 'clinical' factors for short-term risk prediction[77], risk estimation integrating mammographic density and polygenic risk [69,120,121], and contributions to the identification and understanding genomic breast cancer risk by multi-centre consortia[122-125].

The PROCAS 2 study began in 2015 following PROCAS 1, which recruited over 50,000 women eligible for screening mammography at the Great Manchester NHS Breast Screening Programme. Lifestyle, reproductive history and other clinical information was collected via questionnaires, with imaging assessment of mammographic density and DNA obtained for polygenic risk analysis[83]. Numerous studies have been undertaken within the remit of PROCAS, such as the evaluation of Tyrer-Cuzick and Gail risk prediction models in a screening population, the predictive impact of inclusion of mammographic density and/or polygenic risk score components into these models[38,40,84,126,127], extensive assessment of risk feedback and perception[128-130], and probing associations between ethnicity and mammographic density. PROCAS outputs have been central to supporting the feasibility of population-based breast cancer risk assessment, including identifying the lack of major psychological harms of providing 10-year risk estimates from different forms of risk algorithms[130].

Prospective evaluations and trials of risk-based screening

Randomised clinical trials should ideally be used to evaluate risk-stratified screening, with mortality outcomes, balance of harms and benefits, acceptability and cost-effectiveness key to consider. Several key studies are underway[129,131-133], including NCT04359420[129]. This is a non-randomised, counterbalanced study across seven UK screening sites. Women invited to the NHS Breast Screening Programme will either be offered the standard programme, or the additional invitation to use BC-Predict, an automated system offering breast cancer risk assessment (to include questionnaires, breast density measurement and polygenic risk) on invitation to screen[129]. Its aims include assessing risk assessment uptake after offer, uptake of risk consultation, chemoprevention or additional mammography, as well as risks of potential cancer worry, anxiety and health service costs[129].

Some other studies do not seek to evaluate the outcomes of screening intensity/eligibility decisions based on individualised risk estimation. Indeed, some are also exploring how best to communicate personal risk (e.g. PROSPR/PCIPS 3, NCT01879189), promote breast screening uptake based on risk factor-specific educational materials (e.g. NCT00416975), or identify optimal imaging modalities for women at specific levels of predicted risk (e.g. NCT00003736).

The Women Informed to Screen Depending on Measures of risk (WISDOM) trial is a preference-tolerant randomised trial of a risk-based screening algorithm versus standard screening practice in the United States that commenced in 2016[131]. Absolute breast cancer risk estimates are generated using the Breast Cancer Surveillance Consortium (BCSC) model[68], with the integration of a polygenic risk score incorporating 96 SNP loci[131], as well as targeted testing for nine moderate-or-high risk genes (*BRCA1, BRCA2, TP53,*

STK11, PTEN, CHD1, ATM, PALB2 and *CHEK2*). Screening strategies are to be dictated by 5-year predicted risk of incident breast cancer. Regarding women aged 40-49 years: women with a 5-year risk of <1.3% are not being offered screening, those with 5-year risk of 1.3% or greater are undergoing biennial mammography, whilst women with extremely dense breasts or who are carriers of *ATM/PALB2/CHEK2* mutations without a positive family history are undergoing annual mammography. For women aged 50-74 years, all are undergoing biennial mammography unless they are carriers of *ATM/PALB2/CHEK2* mutations without a positive family history in which case they receive annual mammography. Regardless of age group, annual mammography with adjunct MR imaging is being deployed in carriers of *BRCA1/BRCA2/TP53/PTEN/STK11/CDH1* mutations regardless of family history, carriers of *ATM/PALB2/CHEK2* mutations with a positive family history, those that had chest irradiation between the ages of 10-30 or have a 5-year breast cancer risk of at least 6%. With a target recruitment of 100,000 women, it has been projected that 75% of women aged 40-49 years will be allocated to 'no screening' whereas 91% of women aged 50-74 will undergo biennial mammography[131]. The primary endpoints are non-inferiority to standard screening regarding the number of late-stage breast cancers diagnosed (>stage IIB), rates of recall and breast biopsy. Secondary endpoints include the rate of stage IIB and interval cancers, recall rates, rates of ductal carcinoma in situ diagnosis, rates of chemoprevention use, cancer incidence rate, PROMIS anxiety score, and rates of systemic therapy use between arms. Importantly, the design is inherently adaptive so that risk assessment methodology and screening strategies are adjustable in line with future evidence under a 'continuous improvement' framework[134].

The population-based Tailored Breast Screening Trial (TBST, NCT02619123) was initiated in Italy in 2013, and is randomising pre-menopausal women aged 44 years and

older to invitation to 'tailored screening' or an active comparator[133]. The target recruitment is 33,200 women. In the tailored arm, those with BI-RADS grade C-D breast density receive annual mammogram invitations until age 50 and then standard population screening; those with lower density breasts are invited 2-yearly until age 50 then standard population screening. In the active comparator arm, women are invited to annual mammography until age 50, followed by usual population screening. The primary outcome measures are the difference in cumulative interval cancers between arms (also by density group) and the cumulative incidence of >T2 or node positive breast cancers by arms (also by density group). Secondary endpoints include comparison of false positive rates between arms, cumulative incidence of all breast cancer cases and attendance to screening. However, the ramifications of this trial on clinical practice may be limited, due to the basis for stratification (dense versus non-dense breasts), the small divergences in screening strategy (annual or biennial mammography) and the short period in which the screening intensity will be altered.

Another key trial is My Personalized Breast Screening (MyPEBS, NCT03672331); an international study seeking to recruit 85,000 women aged 40-70 in which screening strategy in the experimental arm will be dictated by risk assessment incorporating age, family history, previous benign breast disease, hormone/reproductive history and a polygenic risk score. Specifically, women with one or no first-degree relatives with breast or ovarian cancer will utilise the MammoRisk ® model, otherwise, the Tyrer-Cuzick model will be used (see above). No data regarding the proprietary MammoRisk ® algorithm structure itself, or results of performance evaluation is accessible on the owning company's website (https://www.predilife.com/en/home-2/) or identifiable on Medline, although studies of acceptability/ease of software use are available[135,136]. In the comparator arm, women will be screened with mammography, tomosynthesis or

magnetic resonance imaging (MRI) in accordance with extant national guidelines, whereas in the active arm, estimated 5-year risk will inform mammography and/or tomosynthesis screening every 1 to 4 years (with or without ultrasound depending on breast density). The trial has a non-inferiority design, and the primary outcome is the incidence of stage >II breast cancers for the risk-stratified arm. Secondary outcome measures include a superiority analysis regarding incidence rate of stage >II cancers, rates of false positives and benign biopsies, subject anxiety, health-related quality of life according to the EQ-5D questionnaire[137], and cumulative breast cancer diagnosis rates.

Overall, three key randomised trials are underway to assess the outcomes associated with different screening intervals, starting age, or imaging modalities based on individualised risk assessment, although there is diversity in the robustness of the risk assessment mechanism. The initial results from these trials are likely to emerge within the next 3 years, and it is notable that the largest and most comprehensive is adaptable[131], in that newer methods of risk assessment or amended screening strategies can be incorporated should novel evidence emerge.

Health economic evaluations of stratified screening

If randomised trial evidence did support the efficacy of risk-stratified screening, it is also imperative to understand the health economic impacts of novel approaches. In the context of limited healthcare system resource, potential increased initial costs at the individual level of stratified screening (e.g. genetic testing, algorithm implementation), and that low risk women could be deselected from screening but still develop cancer, cost-

effectiveness studies can not only provide implementation-critical evidence, but also simulate clinical and economic outcomes attainable with different strategies.

Economic simulations of risk-stratified breast screening have modelled the clinical outcomes and cost-effectiveness of a range of scenarios in health systems such as the United Kingdom[138,139], the United States[52,140], Germany[141,142], the Netherlands[143] and China[144]. Health economic modelling thus far has typically evaluated the stratification of screening intensity based on classical risk factors such as positive family history, breast density categories, age, or relative risk based on polygenic risk scores. Outcomes explored include numbers of breast cancer deaths avoided across different strategies, overdiagnoses, and incremental cost-effectiveness ratios (ICERs).

Table 1.4 summarises the outputs of key health economic modelling undertaken in recent years. Generally, these outputs support the overall concept that altering screening practices based on individual risks could offer a more favourable balance of benefits (reduction in breast cancer deaths) and harms (overdiagnosis and unnecessary treatment), or be a more cost-effective approach. However, a cohesive, single narrative regarding a particular algorithm-informed strategy is difficult to synthesise from these results. This is due, in part, to divergence from real-world practice in risk assessment (e.g. simplistic categories of relative risk, or not stating how relative risk would best be assessed), and the lack in some studies of reporting the risk distributions in the target population. Multivariable risk estimation is generally sought, and offers more nuanced assessment than stratifying only on positive versus negative family history, for example. Methods and thresholds for defining strata for risk-stratified screening differ across studies, as do their intentions. Some seek to find the economically optimal screening interval or starting age for screening in cohorts simulated as having set risk levels, or even have intensified screening with additional imaging modalities as one pathway in their models. Few

compare existing age-based methods with risk-stratified approaches in the same target population, and fewer evaluate a more robust view of risk-adapted screening, namely not offering screening to those at lowest risk. Furthermore, the direct comparator is some studies is 'no screening' rather than an evaluation of transitioning from age-based screening to truly risk-adapted screening.

Qualitative research

To accept and engage in screening strategies tailored to individual risk, women need to be able to access and comprehend accurate risk estimations[145-147]. Many women are interested to understand and discuss their risk[116,128,148-150]. Risk communication and risk perception are multifaceted and complex[146], yet it is striking that as few as 10% of women have accurate perceptions of personal risk with an otherwise roughly even split between under- and over-estimators[151,152]. The variable use of absolute and relative risks can have major effects on screening intentions or in some cases be misleading[153]. Absolute risk refers to the probability that an event will occur (e.g. 10% chance that event *X* will happen within time *Y*), whereas relative risks are a comparison between two risks (e.g. a two-fold increase of event *X*). Whilst they may both have utility in discussing risk, conflation of the two can occur and providing different types of risk estimates can be persuasive as well as purely 'informative', mandating clear communication and contextualisation for optimal clinician and patient understanding[145].

A growing body of qualitative research has synthesised evidence from focus groups, semi-structured interviews and other methodologies regarding stakeholder views on the

implementation and acceptability of risk-stratified screening[116,117,154-163]. Particularly key considerations are those pertaining to communication of risk and risk-based pathways across languages and cultural groups[155], socioeconomic groups and those that have lower engagement with preventive healthcare. Generally, perceptions of risk-based screening appear to be favourable, whether based on genetic risk[161,164] or other factors. Whilst felt to be acceptable in principle by many women, the evidence base used to support these approaches needs clear articulation to secure buy-in[154] and reasons for heterogeneous policies for different groups need to be transparent[154]. Importantly, perceptions of risk-stratification as a euphemism for service funding reductions may arise[156], and there should be cognisance of anxiety around self-directed risk assessment if used such as via websites or apps[165] (albeit, at low levels) [162].

Study reference	Population modelled	Modelling approach utilised	Risk groups and screening scenarios simulated	Key results
Sankatsing, et al., 2020 [143]	Netherlands – women without BRCA1/2 mutation Women born in 1974 Time horizon age 40-death	Microsimulation (MISCAN – microsimulation screening analysis; semi-Markov processes)	Risk groups: low, average and high, using "common risk factors", excluding breast density Simulated: biennial screening aged 50-74 overall; biennial or triennial screening for low risk women starting 50-60years to 64-74years; annual or biennial screening for high-risk women starting 40-50 until 74-84 years Assumption of 100% screening attendance	Per 1,000 women: *Biennial screening to all, 50-74:* 206 life-years gained, 16 deaths avoided, 187 false positives, 5 overdiagnosed cases *Triennial screening 50-71 for low-risk:* 134 life years gained, 10 breast cancer deaths avoided, 102 false positives, 3 overdiagnosed cases *Biennial screening 40-74 for high risk:* 380 life years gained, 26 breast cancer deaths avoided, 371 false positives, 7 overdiagnosed cases
Arnold, et al., 2019 [141]	Germany Women aged 50, followed to age 100 or death	Microsimulation Markov model	Risk factors: family history, personal history of biopsy, breast density Compared annual, biennial and triennial universal screening to risk-adapted strategies based on relative risk (3 risk categories) Assumption of 54% adherence	Risk-stratified programmes may be more efficient, depending on mortality reduction or QALYs At 54% adherence, compared with no screening, screening women with relative risk >1 was projected to generate 8.63% mortality reduction, incremental QALYs of 0.023 and incremental costs of 211 Euros per woman (2017 prices).
Sun, et al., 2018 [144]	China (urban population)	Prior natural history Markov model	High-risk defined: relative risk>2 High-risk women: screened using USS aged 40-44 with subsequent mammography if indicated; with both modalities if 45-69 years Low-risk women: no screening (diagnosis after symptoms arise)	*Compared with no screening, risk-adapted approach screening every 3 years, with full treatment:* Lifetime costs US$184 per case (2014 prices), 22.99 QALYs, 0.0127 difference in QALY *Compared with no screening, annual screening and full treatment:*

			Simulated complete treatment and 70% treatment after diagnosis	Lifetime costs US$335.43 per case (2014 prices), 23.01 QALYs, 0.028 different in QALYs
Pashayan, et al., 2018 [138]	UK Cohort n=364,500 Women aged 50 years followed up to 85 years	Life-table model	Three cohorts with screening based on risk group: 1) No screening 2) Screen all women aged 50-69 as per NHS BSP 3) Only women above risk polygenic risk threshold screened every 3 years from age 50-69	*Compared to no screening, risk-based screening:* Overdiagnosis:deaths prevented ratio increased from 0.07 to 0.99 as risk threshold lowered (from 99th to 71st percentile) Minimum ICER at 77th percentile of risk threshold (£11,911 per QALY gained), versus £66,445 when using 99th percentile as risk cut-off (price date not specified) At 32nd percentile of risk, risk-adapted screening generated an incremental cost of £20,066 (price date not specified), 450 more QALYs, and 7 fewer breast cancer deaths
Gray, et al., 2017 [139]	UK Women eligible for NHS BSP (aged 50-70)	Discrete event simulation	Four stratification methods in NHS BSP: 1) Absolute 10-year risk: <3.5% triennial screening, 3.5-8% biennial screening, >8% annual screening 2) Relative 10-year risk: low tertile = triennial, middle tertile = biennial, high tertile = annual 3) Supplemental USS for women with high breast density 4) Approach 1 plus supplemental USS as in (3)	*Compared to current NHS BSP strategy (screening 50-70years of age ever 3 years):* Risk-stratification methods 1 and 2 were deemed cost-effective relative to threshold range of £20,000-30,000 per QALY ICER for method 1 vs. UK BSP: £16,689 (2015 prices)

				ICER for method 2 vs. UK BSP: £23,924 (2015 prices)
Trentham-Dietz, et al., 2016 [52]	US Women aged 50+ Lifetime horizon	Microsimulation models x3	Examined various combinations of breast density (4 categories) and relative risk for factors other than density (1.0, 1.3, 2.0, 4.0) Settings of annual/biennial/triennial screening for women aged 50-74 years, and also for women 65-74 years Assumed 100% adherence	*Per 1000 women with fatty breasts/scattered fibroglandular density + RR 1 or 1.3:* Biennial screening (50-74): 5.1 deaths averted Triennial screening (50-74): 3.4 death averted Biennial screening (50-74): 4.1 deaths averted Triennial screening (65-74): 6.5 deaths averted Triennial screening for average-risk women with low-density breasts provided favourable balance of harms and benefits and is cost-effective Annual screening for higher risk (RR 2.0 or 4.0) with heterogeneously or very dense breasts has favourable balance of benefits and harms and is cost-effective
Schousboe, et al., 2011 [140]	US Women aged 40-49, 50-59, 60-69 and 70-79 (initial mammography at 40) Lifetime horizon	Markov cost-utility microsimulation model	Modelled risk based on BI-RADS breast density category, and up to 2 risk factors (family history or previous biopsy) Examined annual, biennial, triennial, 3-4 yearly mammography or no mammography	A range of cost-effective strategies for women of different age groups, breast density and presence of up to 2 risk factors were identified (assuming $100,000 and $50,000 cost-effectiveness thresholds), for example, at a 50,000 cost-effectiveness threshold: BI-RADS B-D, or BI-RADS A + 1-2 risk factors: biennial screening 50-59 years, reassess at age 60

	BI-RADS A + 0/1 risk factors: 3-4 yearly mammography 50-59 years, reassess at age 60

Table 1.4. Summary of health economic and outcomes models evaluating risk-stratified breast screening identified during the scoping review. BI-RADS = Breast Imaging – Reporting and Data system classification, USS = ultrasound scan, NHS BSP = National Health Service Breast Screening Program (in United Kingdom), ICER = incremental cost-effectiveness ratio, QALY = quality-adjusted life year, RR = relative risk.

Discussion and motivation for the thesis

Overdiagnosis and overtreatment have been widely acknowledged in the prostate cancer screening literature for several years[166,167], and the progression towards risk-stratification of screening or risk-adapted management of diagnosed tumours appears more mature than the field of breast cancer[168-170]. There has been increasing polarisation of the debate regarding screening mammography's benefits and harms over recent decades, manifesting as consistent and at times vociferous disagreement over the interpretation of decades-old trials, the reliability of specific randomised or epidemiological studies, the statistical approaches used to interpret them or the interpretation of epidemiological studies[9,16,19,171-174].

Whilst the prevailing consensus is that screening reduces breast cancer mortality[5], whether this can be improved upon is an avenue of active exploration[175]. This thesis is motivated by the following questions:

1) Is the correct risk trajectory being modelled?

   Not every breast cancer will cause death. Some cancers are detected and treated for no benefit ('overdiagnosed'). This becomes increasingly pertinent as treatments and expected prognosis improve over time (e.g. a 26% reduction in mortality rate is projected from 2014 to 2035[2]). Sub-types vary in their aggressiveness and their proclivity to be screen detected[4]. The risks of being diagnosed with a cancer and the prognosis of that cancer may not be related linearly – indeed, a recent study found weak or even inverse associations between the risks from three incident breast cancer prediction models and prognosis[176].

Lastly, evidence from two leading breast cancer chemoprevention trials (IBIS-I and IBIS-II) showed differential effects against tumour sub-types, and an unclear, i.e. non-significant effect on reducing mortality[177,178]. If the aim of a stratified paradigm is to reduce the numbers of women dying from breast cancer, why rely on increasing an 'interim diagnosis' which may be problematic, rather than structuring screening based on women's risks of mortality?

2) Which modelling approaches can yield the best predictive tool?

There is increasing interest in, and concerns raised regarding the scope for machine learning approaches in clinical prediction modelling to guide stratified medicine[91,99,100,179-181]. Whilst these flexible approaches may be able to capture complex non-linearities and higher-order interaction terms without explicit programming or process some data modalities that standard approaches may not, no single method (statistical or otherwise) represents a panacea in relatively low-dimensional clinical settings with higher signal-to-noise ratios. Superiority of any modelling method based on notions of their ability to capture complex relationships with the outcome assumes that these exist in the source data. Some have criticised the transparency[100], interpretability, risk of algorithmic bias exacerbating extant health inequities[103,182], quality of reporting[180], and ability to handle rare events or censoring of some machine learning approaches[94], as well as the appropriateness of comparisons between regression-based and machine-learning-based models[96]. Given a set prediction task, there is no a priori way to know which approach could yield the 'best' prediction tool. The focus of some papers can be solely on obtaining the highest discrimination index, whilst ignoring calibration and clinical utility. What constitutes the 'best performing'

model has no uniformly agreed definition, but one posits that this should be cognisant of several metrics, model heterogeneity (e.g. stability across relevant sub-groups), and assign large weight to consideration of net benefit approaches which not only implicitly assess 'statistical' measures of performance, but estimate a model's effects on (hopefully) improving clinical decisions.

3) What modalities of information are needed to stratify women, and in what setting?

Studies evaluating the effects of integrating multi-modal data forms into clinical prediction models are typically conducted using data from screening-age or screening-participating women[38,83,183]. Women at unrecognised high risk, but who are too young to be screened[106,107,184] could benefit from expanded access but would not have mammographic data available to feed into a clinical prediction model. Almost half of all breast cancer deaths occur in women older than screening invitation age[2], and some could benefit from screening or prevention strategies in older age if the competing risk of non-breast cancer death is considered appropriately in the context of life expectancy. Therefore, restricting risk-stratified screening studies to those currently eligible for screening does not explore the full scope of the concept's potential – whilst incremental benefits may be seen with multi-model data integration, this data will not be available for many women in a broader potentially eligible population. The relative sufficiency of clinical, polygenic, mammographic and other data forms to accurately stratify the adult female population is uncertain. The setting in which a population stratification mechanism is implemented is also important to consider in terms of the data that will be available to it – such as embedding an algorithm in primary

care software (which may or not be linked to imaging data held elsewhere in the healthcare system).

4) Which risk-stratified strategies could be effective?

   Related to the previous discussion, risk-stratified breast strategies need not be restricted to current practices, i.e. screening ages, or intervals. Should clinical prediction models perform well enough to stratify the adult female population in terms of breast cancer risk (cognisant of the first and second points), what thresholds could be used to inform access to screening services, and what would 'good screening' or 'good prevention' entail in different groups?

This thesis was conceived with the intention to contribute towards reducing breast cancer mortality. Whilst initially conceptualised as one focussed on breast cancer screening, it is apparent that there are parallels between these four motivations with breast cancer prevention, and risk-stratification of breast cancers once they are diagnosed. The scope of this thesis reflects this, and therefore comprises two work packages under the unified aim to develop and evaluate clinical prediction models to inform risk stratification in breast cancer screening, prevention, and management.

The first work package (**Chapters 2 to 5**) comprises exploring a range of regression and machine learning approaches to building and evaluating clinical prediction models using large-scale electronic healthcare record data for three risk trajectories relevant to breast cancer screening, prevention, and management:

1) The 10-year risk of incident breast cancer diagnosis

2) The 10-year 'combined risk' of developing and then dying from breast cancer (in women without cancer at baseline)
3) The 10-year risk of dying from breast cancer after diagnosis

The first and last endpoints have been modelled previously, but existing publications included in systematic reviews may be at high risk of bias and have other methodological limitations precluding implementation[79,81,185,186]. The second endpoint is novel, with no existing models identified in the literature.

In the second work package (**Chapters 6 & 7**), the first two endpoints are focussed on (in the interests of scope and data availability) for an exploration of the incremental effects of integrating polygenic risk estimates and reproductive factor information on the performance of the previously developed clinical prediction tools in screening-age women. This uses the UK Biobank cohort.

The thesis then concludes (**Chapter 8**) with a discussion of the results in the context of the pathway from ideation to model implementation, methodological and ethical considerations that arose during the thesis, and outlines avenues for further research that builds on the work presented here.

The research objectives of this thesis are as follows:

1) For each endpoint (1-3), develop and comparatively evaluate clinical prediction models with different techniques and identify which are the most promising (based on clinical utility analyses and consideration of model performance heterogeneity across relevant sub-groups)
2) For endpoints 1 & 2, assess the incremental effects of integrating polygenic risk and reproductive factor information into the models compared to using routinely collected data available in the primary care electronic healthcare record

Chapter 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**(3): 209-49.

2. Cancer Research UK. Breast cancer diagnosis and treatment statistics. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer. Accessed 27th Nov 2020.

3. Balmana J, Diez O, Rubio IT, et al. BRCA in breast cancer: ESMO Clinical Practice Guidelines. *Ann Oncol* 2011; **22 Suppl 6**: vi31-4.

4. Harbeck N, Penault-Llorca F, Cortes J, et al. Breast cancer. *Nat Rev Dis Primers* 2019; **5**(1): 66.

5. Marmot MG, Altman DG, Cameron DA, et al. The benefits and harms of breast cancer screening: an independent review. *Br J Cancer* 2013; **108**(11): 2205-40.

6. Autier P, Boniol M. Mammography screening: A major issue in medicine. *Eur J Cancer* 2018; **90**: 34-62.

7. Bleyer A, Welch HG. Effect of three decades of screening mammography on breast-cancer incidence. *N Engl J Med* 2012; **367**(21): 1998-2005.

8. Kalager M, Adami HO, Bretthauer M, et al. Overdiagnosis of invasive breast cancer due to mammography screening: results from the Norwegian screening program. *Ann Intern Med* 2012; **156**(7): 491-9.

9. Loberg M, Lousdal ML, Bretthauer M, et al. Benefits and harms of mammography screening. *Breast Cancer Res* 2015; **17**: 63.

10. Brawley OW. Risk-based mammography screening: an effort to maximize the benefits and minimize the harms. *Ann Intern Med* 2012; **156**(9): 662-3.

11. Gotzsche PC, Jorgensen KJ. Screening for breast cancer with mammography. *Cochrane Database Syst Rev* 2013; (6): CD001877.

12. Raichand S, Dunn AG, Ong MS, et al. Conclusions in systematic reviews of mammography for breast cancer screening and associations with review design and author characteristics. *Syst Rev* 2017; **6**(1): 105.

13. Baker SG, Prorok PC. Breast cancer overdiagnosis in stop-screen trials: More uncertainty than previously reported. *J Med Screen* 2020: 969141320950784.

14. Broeders M, Paci E. The balance sheet of benefits and harms of breast cancer population-based screening in Europe: outcome research, practice and future challenges. *Womens Health (Lond)* 2015; **11**(6): 883-90.

15. Jorgensen KJ, Gotzsche PC, Kalager M, et al. Breast Cancer Screening in Denmark: A Cohort Study of Tumor Size and Overdiagnosis. *Ann Intern Med* 2017; **166**(5): 313-23.

16. Monticciolo DL, Helvie MA, Hendrick RE. Current Issues in the Overdiagnosis and Overtreatment of Breast Cancer. *AJR Am J Roentgenol* 2018; **210**(2): 285-91.

17. Zahl PH, Jorgensen KJ, Gotzsche PC. Lead-time models should not be used to estimate overdiagnosis in cancer screening. *J Gen Intern Med* 2014; **29**(9): 1283-6.

18. Zahl PH, Jorgensen KJ, Gotzsche PC. Overestimated lead times in cancer screening has led to substantial underestimation of overdiagnosis. *Br J Cancer* 2013; **109**(7): 2014-9.

19. Marmot MG. Sorting through the arguments on breast screening. *JAMA* 2013; **309**(24): 2553-4.

20. Forrest P. Breast cancer screening. Report to the Health Ministers of England Wales Scotland and N Ireland by a working group chaired by Professor Sir Patrick Forrest. HMSO, 1986.

21. Tabar L, Fagerberg CJ, Gad A, et al. Reduction in mortality from breast cancer after mass screening with mammography. Randomised trial from the Breast Cancer Screening Working Group of the Swedish National Board of Health and Welfare. *Lancet* 1985; **1**(8433): 829-32.

22. Shapiro S, Venet W, Strax P, et al. Ten- to fourteen-year effect of screening on breast cancer mortality. *J Natl Cancer Inst* 1982; **69**(2): 349-55.

23. Richards M. Independent Review of Adult Screening Programmes in England, 2019.

24. Pharoah PD, Sewell B, Fitzsimmons D, et al. Cost effectiveness of the NHS breast screening programme: life table model. *BMJ* 2013; **346**: f2618.

25. Harkness EF, Astley SM, Evans DG. Risk-based breast cancer screening strategies in women. *Best Pract Res Clin Obstet Gynaecol* 2020; **65**: 3-17.

26. Bitencourt AG, Rossi Saccarelli C, Kuhl C, et al. Breast cancer screening in average-risk women: towards personalized screening. *Br J Radiol* 2019; **92**(1103): 20190660.

27. 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**(11): 687-705.

28. Department of Halth & Social Care, NHS England. NHS breast screening (BSP) programme. *https://wwwgovuk/health-and-social-care/population-screening-programmes-breast* Accessed 27th November 2020.

29. United States Preventive Service Task Force. Breast Cancer: Screening. *https://wwwuspreventiveservicestaskforceorg/uspstf/recommendation/breast-cancer-screening* Accessed 27th Nov 2020.

30. Klarenbach S, Sims-Jones N, Lewin G, et al. Recommendations on screening for breast cancer in women aged 40-74 years who are not at increased risk for breast cancer. *CMAJ* 2018; **190**(49): E1441-E51.

31. Rijksinstituut voor Volksgezonheid en Milieu. Bevolkingsonderzoek bortskanker. *https://wwwrivmnl/bevolkingsonderzoek-borstkanker* Accessed 27th Nov 2020.

32. BreastScreen Australia . http://www.cancerscreening.gov.au/internet/screening/publishing.nsf/Content/breast-screening-1. Accessed 27th Nov 2020.

33. National Health Commission Of The People's Republic Of China. Chinese guidelines for diagnosis and treatment of breast cancer 2018 (English version). *Chin J Cancer Res* 2019; **31**(2): 259-77.

34. Ryan R. 'Cochrane Consumers and Communication Review Group: data synthesis and analysis'. http://cccrg.cochrane.org. June 2013.

35. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. *Stat Med* 2004; **23**(7): 1111-30.

36. 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**(3): e007825.

37. Ming C, Viassolo V, Probst-Hensch N, et al. Machine learning-based lifetime breast cancer risk reclassification compared with the BOADICEA model: impact on screening recommendations. *Br J Cancer* 2020; **123**(5): 860-7.

38. Brentnall AR, Cuzick J, Buist DSM, et al. Long-term Accuracy of Breast Cancer Risk Assessment Combining Classic Risk Factors and Breast Density. *JAMA Oncol* 2018; **4**(9): e180174.

39. 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**(1): 21-34.

40. 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**(4): 476-82.

41. Antoniou AC, Cunningham AP, Peto J, et al. The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. *Br J Cancer* 2008; **98**(8): 1457-66.

42. Antoniou AC, Hardy R, Walker L, et al. Predicting the likelihood of carrying a BRCA1 or BRCA2 mutation: validation of BOADICEA, BRCAPRO, IBIS, Myriad and the Manchester scoring system using data from UK genetics clinics. *J Med Genet* 2008; **45**(7): 425-31.

43. Biswas S, Tankhiwale N, Blackford A, et al. Assessing the added value of breast tumor markers in genetic risk prediction model BRCAPRO. *Breast Cancer Res Treat* 2012; **133**(1): 347.

44. Eccles SA, Aboagye EO, Ali S, et al. Critical research gaps and translational priorities for the successful prevention and treatment of breast cancer. *Breast Cancer Res* 2013; **15**(5): R92.

45. Yanes T, Young MA, Meiser B, et al. Clinical applications of polygenic breast cancer risk: a critical review and perspectives of an emerging field. *Breast Cancer Res* 2020; **22**(1): 21.

46. Mavaddat N, Pharoah PD, Michailidou K, et al. Prediction of breast cancer risk based on profiling with common genetic variants. *J Natl Cancer Inst* 2015; **107**(5).

47. Michailidou K, Beesley J, Lindstrom S, et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat Genet* 2015; **47**(4): 373-80.

48. Michailidou K, Hall P, Gonzalez-Neira A, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* 2013; **45**(4): 353-61, 61e1-2.

49. Michailidou K, Lindstrom S, Dennis J, et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* 2017; **551**(7678): 92-4.

50. Winkel RR, von Euler-Chelpin M, Nielsen M, et al. Mammographic density and structural features can individually and jointly contribute to breast cancer risk assessment in mammography screening: a case-control study. *BMC Cancer* 2016; **16**: 414.

51. Sprague BL, Conant EF, Onega T, et al. Variation in Mammographic Breast Density Assessments Among Radiologists in Clinical Practice: A Multicenter Observational Study. *Ann Intern Med* 2016; **165**(7): 457-64.

52. Trentham-Dietz A, Kerlikowske K, Stout NK, et al. Tailoring Breast Cancer Screening Intervals by Breast Density and Risk for Women Aged 50 Years or Older: Collaborative Modeling of Screening Outcomes. *Ann Intern Med* 2016; **165**(10): 700-12.

53. Ekpo EU, McEntee MF. Measurement of breast density with digital breast tomosynthesis--a systematic review. *Br J Radiol* 2014; **87**(1043): 20140460.

54. Magny SJ, Shikhman R, Keppke AL. Breast Imaging Reporting and Data System. StatPearls. Treasure Island (FL); 2020.

55. Van Calster B, Wynants L, Timmerman D, et al. Predictive analytics in health care: how can we know it works? *J Am Med Inform Assoc* 2019; **26**(12): 1651-4.

56. Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 2016; **69**: 245-7.

57. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014; **35**(29): 1925-31.

58. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010; **21**(1): 128-38.

59. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making* 2006; **26**(6): 565-74.

60. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004; **23**(5): 723-48.

61. Riley RD, Ensor J, Snell KI, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 2016; **353**: i3140.

62. Wynants L, Riley RD, Timmerman D, Van Calster B. Random-effects meta-analysis of the clinical utility of tests and prediction models. *Stat Med* 2018; **37**(12): 2034-52.

63. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* 1989; **81**(24): 1879-86.

64. Gail MH, Costantino JP, Pee D, et al. Projecting individualized absolute invasive breast cancer risk in African American women. *J Natl Cancer Inst* 2007; **99**(23): 1782-92.

65. Tice JA, Cummings SR, Ziv E, et al. Mammographic breast density and the Gail model for breast cancer risk prediction in a screening population. *Breast Cancer Res Treat* 2005; **94**(2): 115-22.

66. Zhang X, Rice M, Tworoger SS, et al. 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* 2018; **15**(9): e1002644.

67. Tice JA, Miglioretti DL, Li CS, et al. Breast Density and Benign Breast Disease: Risk Assessment to Identify Women at High Risk of Breast Cancer. *J Clin Oncol* 2015; **33**(28): 3137-43.

68. Tice JA, Cummings SR, Smith-Bindman R, et al. Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model. *Ann Intern Med* 2008; **148**(5): 337-47.

69. Vachon CM, Pankratz VS, Scott CG, et al. The contributions of breast density and common genetic variation to breast cancer risk. *J Natl Cancer Inst* 2015; **107**(5).

70. Rosner B, Colditz GA. Nurses' health study: log-incidence mathematical model of breast cancer incidence. *J Natl Cancer Inst* 1996; **88**(6): 359-64.

71. Rosner B, Colditz GA, Iglehart JD, et al. Risk prediction models with incomplete data with application to prediction of estrogen receptor-positive breast cancer: prospective data from the Nurses' Health Study. *Breast Cancer Res* 2008; **10**(4): R55.

72. Pal Choudhury P, Wilcox AN, Brook MN, et al. Comparative Validation of Breast Cancer Risk Prediction Models and Projections for Future Risk Stratification. *J Natl Cancer Inst* 2020; **112**(3): 278-85.

73. Park B, Ma SH, Shin A, et al. Korean risk assessment model for breast cancer risk prediction. *PLoS One* 2013; **8**(10): e76736.

74. Abdolell M, Payne JI, Caines J, et al. Assessing breast cancer risk within the general screening population: developing a breast cancer risk model to identify higher risk women at mammographic screening. *Eur Radiol* 2020; **30**(10): 5417-26.

75. Wang Y, Gao Y, Battsend M, et al. Development of a risk assessment tool for projecting individualized probabilities of developing breast cancer for Chinese women. *Tumour Biol* 2014; **35**(11): 10861-9.

76. Ueda K, Tsukuma H, Tanaka H, et al. Estimation of individualized probabilities of developing breast cancer for Japanese women. *Breast Cancer* 2003; **10**(1): 54-62.

77. Eriksson M, Czene K, Pawitan Y, et al. A clinical model for identifying the short-term risk of breast cancer. *Breast Cancer Res* 2017; **19**(1): 29.

78. Barlow WE, White E, Ballard-Barbash R, et al. Prospective breast cancer risk prediction model for women undergoing screening mammography. *J Natl Cancer Inst* 2006; **98**(17): 1204-14.

79. Louro J, Posso M, Hilton Boon M, et al. A systematic review and quality assessment of individualised breast cancer risk prediction models. *Br J Cancer* 2019; **121**(1): 76-85.

80. Jaime Caro J, Eddy DM, Kan H, et al. Questionnaire to assess relevance and credibility of modeling studies for informing health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. *Value Health* 2014; **17**(2): 174-82.

81. Zheng Y, Li J, Wu Z, et al. Risk prediction models for breast cancer: a systematic review. *BMJ Open* 2022; **12**(7): e055398.

82. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Ann Intern Med* 2019; **170**(1): W1-W33.

83. Evans DG, Astley S, Stavrinos P, et al. Improvement in risk prediction, early detection and prevention of breast cancer in the NHS Breast Screening Programme and family history clinics: a dual cohort study. Improvement in risk prediction, early detection and prevention of breast cancer in the NHS Breast Screening Programme and

family history clinics: a dual cohort study. Southampton (UK): NIHR Journals Library UK; 2016.

84. Evans DGR, Harkness EF, Brentnall AR, et al. Breast cancer pathology and stage are better predicted by risk stratification models that include mammographic density and common genetic variants. *Breast Cancer Res Treat* 2019; **176**(1): 141-8.

85. Evans DGR, van Veen EM, Harkness EF, et al. Breast cancer risk stratification in women of screening age: Incremental effects of adding mammographic density, polygenic risk, and a gene panel. *Genet Med* 2022; **24**(7): 1485-94.

86. Vilmun BM, Vejborg I, Lynge E, et al. Impact of adding breast density to breast cancer risk models: A systematic review. *Eur J Radiol* 2020; **127**: 109019.

87. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019; **17**(1): 230.

88. Terry MB, Liao Y, Whittemore AS, et al. 10-year performance of four models of breast cancer risk: a validation study. *Lancet Oncol* 2019; **20**(4): 504-17.

89. Acosta JN, Falcone GJ, Rajpurkar P, et al. Multimodal biomedical AI. *Nat Med* 2022; **28**(9): 1773-84.

90. Rajpurkar P, Chen E, Banerjee O, et al. AI in health and medicine. *Nat Med* 2022; **28**(1): 31-8.

91. Wilkinson J, Arnold KF, Murray EJ, et al. Time to reality check the promises of machine learning-powered precision medicine. *Lancet Digit Health* 2020; **2**(12): e677-e80.

92. Christodoulou E, Ma J, Collins GS, Steyerberg EW, et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019; **110**: 12-22.

93. Gravesteijn BY, Nieboer D, Ercole A, et al. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. *J Clin Epidemiol* 2020; **122**: 95-107.

94. Li Y, Sperrin M, Ashcroft DM, et al. Consistency of variety of machine learning and statistical models in predicting clinical risks of individual patients: longitudinal cohort study using cardiovascular disease as exemplar. *BMJ* 2020; **371**: m3919.

95. Van Calster B, Verbakel JY, Christodoulou E, et al. Statistics versus machine learning: definitions are interesting (but understanding, methodology, and reporting are more important). *J Clin Epidemiol* 2019; **116**: 137-8.

96. Smith H, Sweeting M, Morris T, et al. A scoping methodological review of simulation studies comparing statistical and machine learning approaches to risk prediction for time-to-event data. *Diagn Progn Res* 2022; **6**(1): 10.

97. Andaur Navarro CL, Damen JA, Takada T, et al. Systematic review finds "Spin" practices and poor reporting standards in studies on machine learning-based prediction models. *J Clin Epidemiol* 2023; **158**: 99-110.

98. Haibe-Kains B, Adam GA, Hosny A, et al. Transparency and reproducibility in artificial intelligence. *Nature* 2020; **586**(7829): E14-E6.

99. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020; **368**: m689.

100. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* 2020; **368**: l6927.

101. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 2006; **7**: 91.

102. Ganapathi S, Palmer J, Alderman JE, et al. Tackling bias in AI health datasets through the STANDING Together initiative. *Nat Med* 2022; **28**(11): 2232-3.

103. Seyyed-Kalantari L, Zhang H, McDermott MBA, et al. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med* 2021; **27**(12): 2176-82.

104. Yen AM, Tsau HS, Fann JC, et al. Population-Based Breast Cancer Screening With Risk-Based and Universal Mammography Screening Compared With Clinical Breast Examination: A Propensity Score Analysis of 1 429 890 Taiwanese Women. *JAMA Oncol* 2016; **2**(7): 915-21.

105. Liu FC, Lin HT, Kuo CF, et al. Epidemiology and survival outcome of breast cancer in a nationwide study. *Oncotarget* 2017; **8**(10): 16939-50.

106. Mukama T, Fallah M, Tian Y, et al. Risk-tailored starting age of breast cancer screening based on women's reproductive profile: A nationwide cohort study. *Eur J Cancer* 2020; **124**: 207-13.

107. Mukama T, Kharazmi E, Xing X, et al. Risk-Adapted Starting Age of Screening for Relatives of Patients With Breast Cancer. *JAMA Oncol* 2019; **6**(1): 68-74.

108. Duffy SW, Vulkan D, Cuckle H, et al. Effect of mammographic screening from age 40 years on breast cancer mortality (UK Age trial): final results of a randomised, controlled trial. *Lancet Oncol* 2020; **21**(9): 1165-72.

109. Neal CH, Rahman WT, Joe AI, et al. Harms of Restrictive Risk-Based Mammographic Breast Cancer Screening. *AJR Am J Roentgenol* 2018; **210**(1): 228-34.

110. Feig SA. Personalized Screening for Breast Cancer: A Wolf in Sheep's Clothing? *AJR Am J Roentgenol* 2015; **205**(6): 1365-71.

111. Joe BN, Hayward JH. More Lives Risked with Risk-based versus Age-based Breast Cancer Screening. *Radiology* 2019; **292**(2): 329-30.

112. Lee CS, Ashih H, Sengupta D, et al. Risk-Based Screening Mammography for Women Aged <40: Outcomes From the National Mammography Database. *J Am Coll Radiol* 2020; **17**(3): 368-76.

113. Burnside ES, Trentham-Dietz A, Shafer CM, et al. Age-based versus Risk-based Mammography Screening in Women 40-49 Years Old: A Cross-sectional Study. *Radiology* 2019; **292**(2): 321-8.

114. Price ER, Keedy AW, Gidwaney R, et al. The Potential Impact of Risk-Based Screening Mammography in Women 40-49 Years Old. *AJR Am J Roentgenol* 2015; **205**(6): 1360-4.

115. van den Broek JJ, Schechter CB, van Ravesteyn NT, et al. Personalizing Breast Cancer Screening Based on Polygenic Risk and Family History. *J Natl Cancer Inst* 2020.

116. Rainey L, van der Waal D, Broeders MJM. Dutch women's intended participation in a risk-based breast cancer screening and prevention programme: a survey study identifying preferences, facilitators and barriers. *BMC Cancer* 2020; **20**(1): 965.

117. Rainey L, van der Waal D, Donnelly LS, et al. Women's decision-making regarding risk-stratified breast cancer screening and prevention from the perspective of international healthcare professionals. *PLoS One* 2018; **13**(6): e0197772.

118. Rainey L, van der Waal D, Jervaeus A, et al. European women's perceptions of the implementation and organisation of risk-based breast cancer screening and prevention: a qualitative study. *BMC Cancer* 2020; **20**(1): 247.

119. Gabrielson M, Eriksson M, Hammarstrom M, et al. Cohort Profile: The Karolinska Mammography Project for Risk Prediction of Breast Cancer (KARMA). *Int J Epidemiol* 2017; **46**(6): 1740-1g.

120. Vachon CM, Scott CG, Tamimi RM, et al. Joint association of mammographic density adjusted for age and body mass index and polygenic risk score with breast cancer risk. *Breast Cancer Res* 2019; **21**(1): 68.

121. Vachon CM, van Gils CH, Sellers TA, et al. Mammographic density, breast cancer risk and risk prediction. *Breast Cancer Res* 2007; **9**(6): 217.

122. Ferreira MA, Gamazon ER, Al-Ejeh F, et al. Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer. *Nat Commun* 2019; **10**(1): 1741.

123. Escala-Garcia M, Guo Q, Dork T, et al. Genome-wide association study of germline variants and breast cancer-specific mortality. *Br J Cancer* 2019; **120**(6): 647-57.

124. Garcia-Closas M, Couch FJ, Lindstrom S, et al. Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nat Genet* 2013; **45**(4): 392-8, 8e1-2.

125. Garcia-Closas M, Gail MH, Kelsey KT, et al. Searching for blood DNA methylation markers of breast cancer risk and early detection. *J Natl Cancer Inst* 2013; **105**(10): 678-80.

126. Brentnall AR, Cohn WF, Knaus WA, et al. A Case-Control Study to Add Volumetric or Clinical Mammographic Density into the Tyrer-Cuzick Breast Cancer Risk Model. *J Breast Imaging* 2019; **1**(2): 99-106.

127. Brentnall AR, van Veen EM, Harkness EF, et al. A case-control evaluation of 143 single nucleotide polymorphisms for breast cancer risk stratification with classical factors and mammographic density. *Int J Cancer* 2020; **146**(8): 2122-9.

128. Evans DG, Donnelly LS, Harkness EF, et al. Breast cancer risk feedback to women in the UK NHS breast screening population. *Br J Cancer* 2016; **114**(9): 1045-52.

129. French DP, Astley S, Brentnall AR, et al. What are the benefits and harms of risk stratified screening as part of the NHS breast screening Programme? Study protocol for a multi-site non-randomised comparison of BC-predict versus usual screening (NCT04359420). *BMC Cancer* 2020; **20**(1): 570.

130. French DP, Southworth J, Howell A, et al. Psychological impact of providing women with personalised 10-year breast cancer risk estimates. *Br J Cancer* 2018; **118**(12): 1648-57.

131. Shieh Y, Eklund M, Madlensky L, et al. Breast Cancer Screening in the Precision Medicine Era: Risk-Based Screening in a Population-Based Trial. *J Natl Cancer Inst* 2017; **109**(5).

132. Giordano L, Gallo F, Petracci E, et al. The ANDROMEDA prospective cohort study: predictive value of combined criteria to tailor breast cancer screening and new opportunities from circulating markers: study protocol. *BMC Cancer* 2017; **17**(1): 785.

133. Paci E, Mantellini P, Giorgi Rossi P, et al. [Tailored Breast Screening Trial (TBST)]. *Epidemiol Prev* 2013; **37**(4-5): 317-27.

134. Esserman LJ, Study W, Athena I. The WISDOM Study: breaking the deadlock in the breast cancer screening debate. *NPJ Breast Cancer* 2017; **3**: 34.

135. Uzan C, Ndiaye-Gueye D, Nikpayam M, et al. [First results of a breast cancer risk assessment and management consultation]. *Bull Cancer* 2020; **107**(10): 972-81.

136. Weigert J, Cavanaugh N, Ju T. Evaluating Mammographer Acceptance of MammoRisk Software. *Radiol Technol* 2018; **89**(4): 344-50.

137. Devlin N, Pan T, Kreimeier S, et al. Valuing EQ-5D-Y: the current state of play. *Health Qual Life Outcomes* 2022; **20**(1): 105.

138. Pashayan N, Morris S, Gilbert FJ, et al. Cost-effectiveness and Benefit-to-Harm Ratio of Risk-Stratified Screening for Breast Cancer: A Life-Table Model. *JAMA Oncol* 2018; **4**(11): 1504-10.

139. Gray E, Donten A, Karssemeijer N, et al. Evaluation of a Stratified National Breast Screening Program in the United Kingdom: An Early Model-Based Cost-Effectiveness Analysis. *Value Health* 2017; **20**(8): 1100-9.

140. Schousboe JT, Kerlikowske K, Loh A, et al. Personalizing mammography by breast density and other risk factors for breast cancer: analysis of health benefits and cost-effectiveness. *Ann Intern Med* 2011; **155**(1): 10-20.

141. Arnold M, Pfeifer K, Quante AS. Is risk-stratified breast cancer screening economically efficient in Germany? *PLoS One* 2019; **14**(5): e0217213.

142. Arnold M, Quante AS. Personalized Mammography Screening and Screening Adherence-A Simulation and Economic Evaluation. *Value Health* 2018; **21**(7): 799-808.

143. Sankatsing VDV, van Ravesteyn NT, Heijnsdijk EAM, et al. Risk stratification in breast cancer screening: Cost-effectiveness and harm-benefit ratios for low-risk and high-risk women. *Int J Cancer* 2020; **147**(11): 3059-67.

144. Sun L, Legood R, Sadique Z, et al. Cost-effectiveness of risk-based breast cancer screening programme, China. *Bull World Health Organ* 2018; **96**(8): 568-77.

145. Fagerlin A, Zikmund-Fisher BJ, Ubel PA. Helping patients decide: ten steps to better risk communication. *J Natl Cancer Inst* 2011; **103**(19): 1436-43.

146. de Jonge ET, Vlasselaer J, Van de Putte G, et al. The construct of breast cancer risk perception: need for a better risk communication? *Facts Views Vis Obgyn* 2009; **1**(2): 122-9.

147. Edwards AG, Naik G, Ahmed H, et al. Personalised risk communication for informed decision making about taking screening tests. *Cochrane Database Syst Rev* 2013; **2013**(2): Cd001865.

148. Evans DG, Brentnall AR, Harvie M, et al. Breast cancer risk in young women in the national breast screening programme: implications for applying NICE guidelines for additional screening and chemoprevention. *Cancer Prev Res (Phila)* 2014; **7**(10): 993-1001.

149. Evans DG, Howell A. Can the breast screening appointment be used to provide risk assessment and prevention advice? *Breast Cancer Res* 2015; **17**(1): 84.

150. Yanes T, Kaur R, Meiser B, et al. Women's responses and understanding of polygenic breast cancer risk information. *Fam Cancer* 2020; **19**(4): 297-306.

151. Printz C. Most women have an inaccurate perception of their breast cancer risk. *Cancer* 2014; **120**(3): 314-5.

152. Abittan B, Pachtman S, Herman S, et al. Perception of Breast Cancer Risk in Over 11,000 Patients During Routine Mammography Exam. *J Cancer Educ* 2020; **35**(4): 782-7.

153. Gigerenzer G. Breast cancer screening pamphlets mislead women. *BMJ* 2014; **348**: g2636.

154. McWilliams L, Woof VG, Donnelly LS, et al. Risk stratified breast cancer screening: UK healthcare policy decision-making stakeholders' views on a low-risk breast screening pathway. *BMC Cancer* 2020; **20**(1): 680.

155. Woof VG, Ruane H, French DP, et al. The introduction of risk stratified screening into the NHS breast screening Programme: views from British-Pakistani women. *BMC Cancer* 2020; **20**(1): 452.

156. He X, Schifferdecker KE, Ozanne EM, et al. How Do Women View Risk-Based Mammography Screening? A Qualitative Study. *J Gen Intern Med* 2018; **33**(11): 1905-12.

157. Schifferdecker KE, Tosteson ANA, Kaplan C, et al. Knowledge and Perception of Breast Density, Screening Mammography, and Supplemental Screening: in Search of "Informed". *J Gen Intern Med* 2020; **35**(6): 1654-60.

158. Ghanouni A, Sanderson SC, Pashayan N, et al. Attitudes towards risk-stratified breast cancer screening among women in England: A cross-sectional survey. *J Med Screen* 2020; **27**(3): 138-45.

159. Ghanouni A, Waller J, Stoffel ST, et al. Acceptability of risk-stratified breast screening: Effect of the order of presenting risk and benefit information. *J Med Screen* 2020; **27**(1): 52-6.

160. Lippey J, Keogh LA, Mann GB, et al. "A Natural Progression": Australian Women's Attitudes About an Individualized Breast Screening Model. *Cancer Prev Res (Phila)* 2019; **12**(6): 383-90.

161. Meisel SF, Pashayan N, Rahman B, et al. Adjusting the frequency of mammography screening on the basis of genetic risk: Attitudes among women in the UK. *Breast* 2015; **24**(3): 237-41.

162. van Erkelens A, Sie AS, Manders P, et al. Online self-test identifies women at high familial breast cancer risk in population-based breast cancer screening without inducing anxiety or distress. *Eur J Cancer* 2017; **78**: 45-52.

163. Fürst N, Kiechle M, Strahwald B, et al. Mammography Screening 2.0 - How Can Risk-Adapted Screening be Implemented in Clinical Practice?: Results of a Focus Group Discussion with Experts in the RISIKOLOTSE.DE Project. *Geburtshilfe Frauenheilkd* 2018; **78**(5): 506-11.

164. Wong XY, Chong KJ, van Til JA, et al. A qualitative study on Singaporean women's views towards breast cancer screening and Single Nucleotide Polymorphisms (SNPs) gene testing to guide personalised screening strategies. *BMC Cancer* 2017; **17**(1): 776.

165. Elkin EB, Pocus VH, Mushlin AI, et al. Facilitating informed decisions about breast cancer screening: development and evaluation of a web-based decision aid for women in their 40s. *BMC Med Inform Decis Mak* 2017; **17**(1): 29.

166. Loeb S, Bjurlin MA, Nicholson J, et al. Overdiagnosis and overtreatment of prostate cancer. *Eur Urol* 2014; **65**(6): 1046-55.

167. Vickers AJ, Sjoberg DD, Ulmert D, et al. Empirical estimates of prostate cancer overdiagnosis by age and prostate-specific antigen. *BMC Med* 2014; **12**: 26.

168. Pashayan N, Duffy SW, Neal DE, et al. Implications of polygenic risk-stratified screening for prostate cancer on overdiagnosis. *Genet Med* 2015; **17**(10): 789-95.

169. Vickers AJ. Redesigning Prostate Cancer Screening Strategies to Reduce Overdiagnosis. *Clin Chem* 2019; **65**(1): 39-41.

170. Vickers AJ, Sud A, Bernstein J, et al. Polygenic risk scores to stratify cancer screening should predict mortality not incidence. *NPJ Precis Oncol* 2022; **6**(1): 32.

171. Baum M. Should routine screening by mammography be replaced by a more selective service of risk assessment/risk management? *Womens Health (Lond)* 2010; **6**(1): 71-6.

172. Bleyer A. Were our estimates of overdiagnosis with mammography screening in the United States "based on faulty science"? *Oncologist* 2014; **19**(2): 113-26.

173. Nelson HD. Mammography Screening and Overdiagnosis. *JAMA Oncol* 2016; **2**(2): 261-2.

174. Paci E, Broeders M, Hofvind S, et al. European breast cancer service screening outcomes: a first balance sheet of the benefits and harms. *Cancer Epidemiol Biomarkers Prev* 2014; **23**(7): 1159-63.

175. Kerlikowske K, O'Kane ME, Esserman LJ. Fifty years of age-based screening: time for a new risk-based screening approach. *Evid Based Med* 2014; **19**(5): 183.

176. 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**(8): e0160966.

177. 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**(1): 67-75.

178. 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**(10218): 117-22.

179. Kundu S. AI in medicine must be explainable. *Nat Med* 2021; **27**(8): 1328.

180. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet* 2019; **393**(10181): 1577-9.

181. Van Calster B, Steyerberg EW, Collins GS. Artificial Intelligence Algorithms for Medical Prediction Should Be Nonproprietary and Readily Available. *JAMA Intern Med* 2019; **179**(5): 731.

182. Li J, Bzdok D, Chen J, et al. Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Sci Adv* 2022; **8**(11): eabj1812.

183. Brentnall AR, Harkness EF, Astley SM, et al. Mammographic density adds accuracy to both the Tyrer-Cuzick and Gail breast cancer risk models in a prospective UK screening cohort. *Breast Cancer Res* 2015; **17**(1): 147.

184. Mukama T, Kharazmi E, Sundquist K, et al. Risk-adapted starting age of breast cancer screening in women with a family history of ovarian or other cancers: A nationwide cohort study. *Cancer* 2021; **127**(12): 2091-8.

185. Phung MT, Tin Tin S, Elwood JM. Prognostic models for breast cancer: a systematic review. *BMC Cancer* 2019; **19**(1): 230.

186. Hueting TA, van Maaren MC, Hendriks MP, et al. The majority of 922 prediction models supporting breast cancer decision-making are at high risk of bias. *J Clin Epidemiol* 2022; 152: P238-247.

187. Yang X, Eriksson M, Czene K, et al. Prospective validation of the BOADICEA multifactorial breast cancer risk prediction model in a large prospective cohort study. *J Med Genet* 2022; 59(12): 1196-1205.

188. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med* 2008; 27(20):157-172.

189. Hilden J & Gerds TA. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Stat Med* 2014; 33(19):3405-3414.

190. Pepe MS, Jan J, Feng Z, et al. The net reclassification index (NRI): a misleading measure of prediction impriovement even with independent test data sets. *Stat Biosci* 2015; 7(2): 282-295.

191. Hurson AN, Pal Choudhury PP, Gao C, et al. Prospective evaluation of a breast-cancer risk model integrating classical risk factors and polygenic risk in 15 cohorts from six countries. *Int J Epidemiol* 2021; 50(6): 1897-1911.

192. Ho PJ, Ho WK, Khng AJ, et al. Overlap of high-risk individuals predicted by family history, and genetic and non-genetic breast cancer risk prediction models: implications for risk-stratification. *BMC Medicine* 2022; 20:150.

Chapter Two

Methods for clinical prediction model development and evaluation

Parts of this chapter have been peer-reviewed and published previously elsewhere:

Clift, AK. 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

Summary

This chapter details and explains the rationale for the methodological approaches used for 'work package one', which seeks to develop and evaluate clinical prediction models for three outcomes relevant to breast cancer screening, prevention, and management. This work package is conducted using the QResearch primary care database and its linked data sources.

Introduction

The aphorism that "*all models are wrong, but some are useful*" is optimistic, given that the overwhelming majority of clinical prediction models developed are never used[1,2]. The clinical prediction modelling literature is markedly heterogeneous in quality, the robustness of analytical methods employed, and the clinical usefulness of developed tools. This complexity is furthered by contemporary interest in 'machine learning' approaches to prediction modelling. Some have posited that these flexible algorithms could better capture complex interactions or non-linear associations without explicit programming[3,4]. Others have raised concerns regarding the methodological rigour in many ML studies[5–7], model transparency[5], explainability, interpretability[8], risk of 'algorithmic bias' exacerbating extant health inequalities[9], ability to handle rare events or censoring[10,11], and appropriateness of comparisons to regression-based methods (particularly when reporting novel algorithmic approaches)[12]. For example, some recent papers have compared the average performance of complex ensembles to a simplistic Cox model without any interactions and only a single, pre-determined (i.e. non data-driven non-linear term) term for age[13].

There is no *a priori* way to decide which approach may yield the 'best' model and the evaluation strategy should seek to be as informative as possible[14]. This chapter details the methodological approach to the development and comparative evaluation of several statistical and machine learning models using QResearch, a primary care electronic healthcare record database of routinely collected clinical data, and its linked datasets. As stated in the previous chapter, there is no consensus agreement on what constitutes the 'best' model in studies that compare different approaches – this thesis bases model comparisons in decision curve analysis (as it implicitly integrates discrimination and

calibration with an emphasis on usefulness to clinical decision making) and adds a consideration of model stability across pre-specified sub-groupings.

Methods

Study population and data sources

A cohort study approach was used for all modelling. Cohort studies comprise individuals without an event of interest at baseline, and tracks the status of individuals through time until the end of the study period. These can be prospective (individuals are recruited, baseline data collected, and monitored over time thereafter for the occurrence of the outcome[s] of interest), or retrospective (where baseline and outcome statuses are determined using already collected data). In closed cohort studies, the membership of the cohort is fixed insofar as once the participants are recruited, no further individuals are added. In open cohort studies, individuals can be added over time – for example, women could enter the cohort from an eligible population on reaching a certain age later in the study period. As such, open cohorts can permit larger sample size attainment.

An open cohort of women (self-reported female sex) aged between 20 and 90 years at entry into the QResearch database was identified. The study period was 1st January 2000 to 31st December 2020. Women were eligible to enter the cohort on the latest of: date of their 20th birthday, date of registration with the general practice plus 1 year, or date the practice began contributing to QResearch plus 1 year.

Study participants had to have a recorded anonymised version of their NHS number in the QResearch database to facilitate individual-level linkage of information from the different data sources. QResearch collects information on factors such as demographics, prescriptions, diagnoses, laboratory tests and results, symptoms, appointments/consultations and referrals in primary care. No free text information is used – instead, these variables are defined as the presence of 'clinical codes' that are entered by healthcare practitioners into the electronic health record. QResearch uses two types of clinical coding systems, Read and Systematised Nomenclature of Medicine Clinical Terms (SNOMED). Briefly, the occurrence of a clinical 'event' is defined as the presence of a relevant code for that event in the electronic health record. These groups of codes are specified by the analysis team prior to data extract.

This primary care-derived data had linkage at the individual level to several other databases. First, to NHS Digital's Hospital Episode Statistics (HES) which records information about hospital encounters (e.g. accident & emergency attendances, outpatient appointments, inpatient admissions) – this is based on 'episodes', which have dates recorded, and diagnoses made during these episodes are recorded using the International Classification of Diseases version 10 (ICD-10) codes. Second, to the National Cancer Registration and Analysis Service (NCRAS) which is England's population-based cancer registry – it receives information from NHS England regarding individuals diagnosed with non-malignant and malignant tumours, and collates data submitted from service providers from several sources such as multidisciplinary team meetings, hospital administration systems and the Cancer Outcomes and Services Dataset. It has had national coverage since 1971. Last, to the Office for National Statistics' mortality register, which records information on date and case of death (based on information on death certificates).

Outcome definitions

There were three clinical outcomes of interest – each was modelled separately using data derived from the open cohort described above, and the results for which are detailed in subsequent chapters in this thesis:

1) 10-year risk of incident breast cancer diagnosis (**endpoint 1**)
2) 10-year risk of developing and then dying from breast cancer (**endpoint 2**)
3) 10-year risk of breast cancer mortality following diagnosis of invasive breast cancer (**endpoint 3**)

For endpoint 1, the denominator was all women without a pre-existing diagnosis of breast cancer or ductal carcinoma in situ (DCIS) at date of cohort entry, defined as the presence of relevant clinical codes in either primary care or HES datasets. Follow-up was calculated from cohort entry to recorded breast cancer diagnosis date or censoring (left cohort, reached end of study period alive or died from any non-breast cancer cause). Breast cancer diagnosis was defined as the presence of Read/SNOMED codes in primary care, selected ICD-10 codes in HES, a record in the cancer registry, or as a cause of death (primary or contributory) in the death register. The event date was taken as the earliest of these records in any dataset. Women with DCIS were excluded from the denominator as these non-obligate precursor lesions are typically identified during mammography and present a high-risk group with their own clinical considerations[15].

For endpoint 2, the denominator was also all women without pre-existing recorded diagnoses of breast cancer or DCIS. Follow-up was calculated from cohort entry to date of breast cancer death derived from the ONS mortality register, or censoring (left cohort, reached end of study period alive or died from another cause [competing event]). Breast

cancer death was defined as its presence as the primary or contributory cause of death on the death certificate, i.e. any mention of breast cancer in part I or part II of death certificates. This enabled capture of direct deaths and 'indirect' deaths from the perspective of death certification, such as an intracranial haemorrhage caused by brain metastases.

For endpoint 3, the denominator was women diagnosed with invasive breast cancer during the study period. The dataset for this outcome was derived from the original cohort. Those with DCIS recorded in GP/HES, or 'stage 0' tumours in the cancer registry were excluded. Follow-up was calculated from date of breast cancer diagnosis (earliest date recorded in GP/HES/cancer registry) to the date of breast cancer death (obtained from ONS death register) or censoring (left cohort, reach study end date alive or died from another cause). Breast cancer mortality was defined as per endpoint 2.

Prior to modelling, crude and age group-specific (5 years) rates of incident breast cancer diagnoses and breast cancer mortality were calculated for the study period overall, and for two phases, phase 1 (1 January 2000 to 31 December 2009) and phase 2 (1 January 2010 to 31 December 2020). For endpoint 1 (incident breast cancer risk), the ascertainment yields of breast cancer cases from each of the four linked datasets were compared. Clinicopathological characteristics obtained from linked cancer registry data were tabulated for breast cancer cases, and temporal trends in recording completeness were examined.

For all endpoints, the competing event of non-breast cancer death was considered in the modelling (*see below).*

Candidate predictor parameters

These were informed by published evidence from clinical, epidemiological, risk modelling and pre-clinical/molecular studies regarding each endpoint[16–25], and are summarised in **Table 2.1**. For selected comorbidities that could affect risk, relevant treatments were also considered, with exposure to a medication defined by the presence of at least three recorded prescriptions prior to the cohort entry date. Medication use was a binary categorical variable for all models with the exception of hormone replacement therapy (HRT) – based on previously published evidence regarding differential associations with breast cancer risk by type, recency and duration of HRT prescription, the same multilevel categorisation of Vinogradova, et al. was used[80].

The latest recorded measurement prior to/at cohort entry was considered for inclusion in the modelling with no limit on recency. Clinical diagnoses used as predictor variables were defined as either being recorded in primary care data (Read/SNOMED codes) or in hospital records (ICD-10 codes) on/prior to the cohort entry date.

The clinical code groups used for this thesis are available on the QWeb platform: www.qresearch.org/data/qcode-group-library/. **Appendix 1** lists the QWeb code groups used to define predictors.

Missing data

Given the longitudinal nature of electronic healthcare record data and that recording of some variables may depend on clinical context, there were incomplete data for some variables, namely self-reported ethnicity, smoking status, alcohol intake, and Townsend deprivation score for all 3 endpoints. In QResearch, the Townsend deprivation score is

evaluated at ‘output areas’, which each comprise approximately ten households. For endpoint 3, there was also missing data in several fields obtained from the cancer registry: stage at diagnosis, tumour grade, oestrogen receptor status, progesterone receptor status, and HER2 status.

Handling missing data is an important consideration for clinical prediction modelling (and epidemiological research in general) and can be performed in different ways. The validity of approaches for handling missing data depend on the missingness mechanism[26–30]. Rubin articulated three such mechanisms[27]:

1) *Missing completely at random (MCAR)* – there is no relationship between the values of the missing data and their probability of not being recorded, and the distributions of recorded and non-recorded data are similar
2) *Missing at random (MAR)* – the probability of missingness and the missing values are related to the values of the recorded data, therefore the measured data could be used to recover missing information
3) *Missing not at random (MNAR)* – there is a relationship between the value of the missing data points themselves and their likelihood of being missing

Complete case analyses risk biased coefficients unless incomplete data are MCAR, but this can result in a loss of precision[30]. Single imputation may not capture enough uncertainty regarding around the missing values, and simple measures such as mean/median imputation artificially force predictor distributions away from the ‘true’ towards being less heterogeneous. Under MAR, multiple imputation with chained equations may be a valid approach[26]. This provides a flexible, Bayesian approach where multiple datasets are generated and replacement values for missing data points found by

sampling from the predictive distributions based on the measured data[26]. These multiply imputed datasets can then be analysed separately, and the results from analysing each imputation pooled to provide overall estimates – this can take into account the 'in-imputation' and 'between-imputation' variance to provide appropriate standard errors around point estimates[27].

Identifying the missingness mechanism depends on unobservable data and is therefore unprovable; rather, it is the plausibility of assumptions that should be considered. Through descriptors such as temporal trends in recording, and cross-tabulation of the final datasets, the missing at random assumption was deemed plausible and therefore multiple imputation with chained equations was used in each modelling scenario to handle missing values[31]. This was applied to the overall cohort.

In the final cohort dataset for each endpoint, the imputation models contained all respective candidate predictors, the endpoint indicator, the Nelson-Aalen cumulative hazard estimate[31], and decade of cohort entry as an auxiliary variable. Auxiliary variables are those that are not included in the model, but are correlated with the outcome of interest and can therefore help 'recover' missing information. For endpoint 3, the breast cancer treatments used within the first year after diagnosis, as obtained from the cancer registry, were included as additional auxiliary variables (i.e. mastectomy, other breast surgery, chemotherapy, and radiotherapy). Five imputations were generated for the datasets used to model endpoint 1 and 2 due to considerations of computational resource; fifty imputations were generated for the endpoint 3 dataset due its smaller size and considering the fraction of missing information[32]. Multiply imputed data were used throughout all model fitting and evaluation steps for all models. Recently, there has been some discussion in the literature regarding the mismatch between methods used to handle missing data during model development and those used at point of model use[28,33] – such

mismatch could lead to incorrect estimation of model performance in the real-world prospective implementations. Some pre-print simulation modelling studies suggest that regression imputation or missing value indicators may marginally outperform multiple imputation, but there is no clear best practice[33]. Therefore, this work abided by existing standard principles and used multiple imputation.

Modelling methods

Four methods for modelling were explored for each endpoint:

1) Cox proportional hazards regression
2) Competing risks regression
3) Extreme gradient boosting (XGBoost)
4) Artificial neural networks

Cox proportional hazards regression

The Cox proportional hazards model can summarise the relative effects of covariates on survival by time *t*:

$$S(t) = S_0(t)^{\exp(\beta 1X1 + \cdots + \beta aXa)}$$

Where $S_0(t)$ is the baseline hazard function at time *t*, and β_a is the coefficient for variable X_a. The Cox proportional hazards model is semi-parametric insofar as it does not specify the baseline function. Clinical prediction models can estimate an individual's risk by calculating $1 - S_{(t)}$, with the baseline function estimated when categorical covariates are set to zero and with the mean value of continuous variables.

Predictor class	Variables (and raw functional form)
Demographic variables	Age (continuous variable) Townsend deprivation score (continuous) Ethnicity (categorical, as per Office for National Statistics Census classes) *
Lifestyle factors	Smoking status (categorical; non-smoker, ex-smoker, light smoker [<10/day], moderate smoker [11-19/day], heavy smoker [20+/day]) Body mass index (continuous) Alcohol intake (categorical; <1 unit/day, 1-2units/day, 3-6 units/day, 7-9 units/day, 10+units/day)
Co-morbidities and medical history (all binary, unless otherwise specified)	Previous gynaecological cancer (ovarian/endometrial/uterine) Previous lung cancer Previous haematological cancer (leukaemia, lymphoma or myeloma) Previous thyroid cancer Hypertension Ischaemic heart disease Diabetes mellitus type 1 Diabetes mellitus type 2 Cirrhosis of the liver/chronic liver disease Systemic lupus erythematosus Psychosis (inc. schizophrenia, depression with psychosis) Fibromatosis or fibrocystic breast disease Polycystic ovarian syndrome Endometriosis Chronic kidney disease (ordinal categorical, none/stage 1 or 2, then stages 3-5 [end-stage renal failure]) Vasculitis
Family history	Recorded family history of gynaecological cancer Recorded family of breast cancer
Medications (at least 3 prescriptions prior to cohort entry; binary categorical)	Anti-hypertensives: angiotensin converting enzyme inhibitors, calcium channel blockers, renin angiotensin aldosterone axis antagonists, beta blockers Anti-psychotics (atypical or typical) Tricyclic antidepressants Selective serotonin reuptake inhibitors Monoamine oxidase inhibitors Hormone replacement therapy** Oral contraceptive pill use
Reproductive history	Number of pregnancies (continuous)*** Menopause (binary; *defined as recorded diagnosis of menopause on GP or HES records, recorded prescriptions of hormone replacement therapy, or age at 60 at entry*)
Tumour characteristics (for diagnosed tumours):	Stage at diagnosis (ordinal categorical, I-IV) Tumour grade (differentiation; ordinal categorical) Oestrogen receptor status (binary) Progesterone receptor status (binary) Human epidermal growth factor receptor 2 (HER2) status (binary) Route to diagnosis (e.g. two-week referral, emergency presentation, screen-detected)

Table 2.1. Candidate predictors considered in the clinical prediction models. * = ethnicity classification for each endpoint was informed by event counts in different strata – see later results chapters for when categories were collapsed down on basis of geographical similarity. ** = hormone replacement therapy was classified by type (oestrogen-only or combined), and sub-classified by recency of last prescription (<5 or 5+ years since) and duration of prescription (<1 year, 1-2 years, 3-4 years, 5-9 years, 10+ years). *** = during data processing, the very high extent of data missingness (>90%) and counterintuitive distributions of parity in the general practice data led to a decision to not include number of pregnancies in the modelling.

Competing risks regression

Competing events preclude the occurrence of the endpoint of interest; for example, a woman dying from cardiovascular disease can by definition not then develop an incident breast cancer. In such settings, Cox proportional hazards models can overestimate risk as they inappropriately treat competing events as censoring[34–36]. Rather than using the complement of the Kaplan-Meier survival function to estimate risks, the cumulative incidence function (CIF) of the event of interest is considered[34].

Two general forms of hazard function are relevant in competing risks scenarios – the cause-specific hazard function, and the subdistribution hazard function[37,38]. The cause-specific hazard function expresses the instantaneous rate of the event of interest occurring in individuals that are free from experiencing any event (endpoint or competing):

$$\lambda_k^{cs}(t) = \lim_{\Delta t \to 0} \frac{\text{Prob}(t \le T < t + \Delta t, D = k \mid T \ge t)}{\Delta t}.$$

The subdistribution hazard function expresses the instantaneous risk of the event of interest occurring in those that have not yet experienced it:

$$\lambda_k^{sd}(t) = \lim_{\Delta t \to 0} \frac{\text{Prob}(t < T \le t + \Delta t, D = k \mid T > t \cup (T < t \cap K \ne k))}{\Delta t}.$$

For both of these equations, we assume that there are '*K*' different types of events that may occur, '*T*' denotes the time to failure from the '*k*'th event type, and '*D*' is a variable denoting the type of event that has occurred.

Cause-specific hazards modelling uses separate Cox proportional hazards models for each outcome[35], but it has been suggested that this approach is best suited to studying causation or aetiology, whilst modelling in a subdistribution hazards framework may best for predicting risks since it directly models the CIF for the endpoint of interest[39,40]. The archetypal subdistribution approach is the Fine-Gray model, but this can be difficult to interpret[37] and computationally intensive, reducing feasibility in datasets as large as QResearch. The following alternative was developed from the epidemiological literature to provide a single model approach that permits direct modelling of probabilities [41,42].

Using the non-parametric Aalen-Johansen estimator for the event-specific CIF of cause *i* at time *t*:

$$CIF_{(i)(t)} = \sum_{tj \leq t} \frac{d_{ij}}{Y_j} \hat{S}(t_{j-})$$

Where d_{ij} is the number of events of cause *i* at time t_j, Y_j is the number at risk at time t_j, and $S(t_j-)$ is the left-sided limit of the survival function at time t_j, one can estimate 'pseudo-observations' for each individual using a jack-knife estimator[41,42]:

$$\widehat{\theta_i} = n\hat{\theta} - (n-1)\hat{\theta}_{-1}$$

Here, the pseudo-observation for individual *i* is denoted as $\widehat{\theta_i}$. This is calculated by using *n* (the number of observations) multiplied by $\hat{\theta}$ (the CIF as estimated on the whole sample), then subtracting (*n*-1) times the CIF estimate that is obtained after excluding the

data for individual *i* ($\hat{\theta}_{-1}$). Therefore, for this procedure, the whole sample CIF is estimated once in all *n* individuals, then used to estimate all individuals' pseudo-observations. When calculating each individual's pseudo-observation, the CIF is also re-estimated after holding out the data from individual *i* ($\hat{\theta}_{-1}$).

These can be considered as an individual's contribution to the CIF, act as individual pseudo-observed probabilities for the outcome of interest, and can be used as a continuous outcome variable in a generalised linear model with a complementary log-log link function[41] to approximate a subdistribution hazards model in a computationally tractable way without assuming proportionality of subdistribution hazards. Robust standard errors are used to account for non-independence of pseudo-observations. Exponentiated coefficients from this model could be interpreted similarly to subdistribution hazard ratios, and model predictions can be converted into probabilities using an inverse of the complementary log-log link function:

$$Predicted\ event\ probability = 1 - \exp\left(-\exp\left(\beta_1 X_1 + \cdots + \beta_k X_k\right)\right.$$

The same approach can be used to estimate pseudo-observations for the Kaplan-Meier failure function, which permits direct modelling of failure probabilities in a non-competing risks framework.

Extreme gradient boosting (XGBoost)

Gradient boosting describes the iterative fitting of individually 'weak' learners and combining them to form a collectively stronger model ('ensemble'). Extreme gradient boosting (XGBoost) is one such approach[43].

After initialising a single decision tree-based model with a constant value, the errors between observed and predicted values are computed ('residuals') which then inform the computation of first ('gradients') and second ('hessians') derivatives of a selected loss function, fits another tree that seeks to minimise these errors[43]. The iterative improvement with each tree fitted to minimise the errors from the previous and the magnitude of these corrections are influenced by the learning rate. This is repeated *M* times to form an ensemble model, whose predictions are the sum of estimates from all trees.

XGBoost software packages support tree-based variants of the accelerated failure time model, and the Cox proportional hazards model[44]. The former is not suitable for the risk prediction sought here as it estimates survival times. The latter was not deemed optimal for this thesis as the model returns exponentiated results on the hazard ratio scale (which would need to be combined with a classically derived baseline function), and it does not have graphical processing unit (GPU) support[45] due to the nature of the likelihood being minimised. This is a major limitation as on very large datasets, computation time can be considerable running without GPU support, particularly within the planned analysis and validation framework that includes hyperparameter tuning.

Instead, pseudo-observations approaches were also used for all XGBoost models. Using the continuous pseudo-observations as the outcome with a squared error objective (loss) function and a root mean squared error evaluation metric (both have GPU support), the

time-to-event task (potentially with competing risks) is converted to a regression task in a way that retains the potential benefits of XGBoost's flexibility.

Artificial neural networks

Feedforward artificial neural networks (ANNs) comprise interconnected layers of functions ('nodes') that transform input data into an output prediction via successive abstractions[3,46]. They comprise a set of flexible approaches that can be used to model structured ('tabular'), or non-tabular data, e.g. in natural language processing or image analysis[3,47].

ANNs are generally parameterised by assigning weights to the interconnections between nodes, activation functions that transform the 'input' to each node (e.g. from the previous layer), bias terms (multiplicative terms on the activation function), and activation functions in the output layer node(s) generates the final prediction(s). Fitting a neural network involves identification of the optimal values for its parameters by defining a loss function (distance between predictions and true values), and using gradient descent (updating weights in the opposite direction to the gradient of the function at the given values) to minimise this. After passing batches of data through the network and generating predictions, iterative changes to the model parameters are informed by 'backpropagation'. Moving backwards towards the input layer, this estimates the gradient of the loss function with respect to each weight in each layer, informing weight updates.

Neural network adaptations of the Cox model have been developed, such as DeepSurv, a feedforward network that predicts an individual's hazard by using a regularised version of the Cox model's negative partial likelihood as a loss function[48]. Similar to the XGBoost modelling, this thesis used a pseudo-observation method to model risk probabilities

directly, potentially in the setting of competing risks, rather than using methods that output a value that need to be combined with a classically estimated baseline function. Such approaches have been reported previously but do not seem to have had wide adoption[49].

Feedforward architectures are commonly used in tabular, low-dimensional medical data[50–53], and were selected for this thesis due to their relative simplicity whilst retaining the possibility to model complex interactions or capture complex functional forms with densely connected hidden layers. Other approaches such as 'ResNet' perform well with non-tabular data settings, such as in image analysis.

Model development and evaluation strategy – overall

The approach used to estimate performance metrics can implicitly affect their interpretation. The archetypal 'split sample' approach wherein the source dataset is randomly partitioned is attractive in its simplicity and provision of 'unseen data' to estimate model performance. However, in small samples, random splitting is statistically inefficient, wastes data, and is therefore inadvisable[54]. In very large samples, random splitting likely yields two datasets that are similar in terms of predictor distributions and outcomes. This is essentially test of reproducibility by testing on data drawn from the same underlying population[55], rather than estimating how well the model might transport or generalise to 'new' and potentially different populations.

Rather than asking "does this model work on unseen data?" model evaluation can be framed as: "having developed this model, how might it work when applied to new individuals?" A prospectively implemented model could be used in geographically distinct regions, and there may be temporal trends in the incidence of the event of interest,

hence the definition of external validation encompassing evaluation in geographically or temporally distinct samples[56,57]. Both can affect performance, the latter being acutely demonstrated in the need for temporal recalibration of COVID-19 prediction models[58–60], but may also be the case for breast cancer incidence and mortality[61]. Terms such as 'clinician and dataset shift' or 'calibration drift' have been used to describe this[62].

For all three endpoints, models were fitted to the entire available data, and an internal-external cross-validation (IECV) framework was used to estimate their performance and both geographical and temporal transportability[63] (**Figure 2.1**). This IECV approach used the long duration of the open cohort (20 years; models have a 10-year prediction horizon) derived from QResearch, and used non-random splitting by geographical region and time period. This iteratively refits the model using data from all-but-one region in the first decade, estimates the performance using the data from the held-out region from the later time period, and repeats for each region. Thus, it provides information on performance heterogeneity, and estimates how well models may generalise to new individuals/transport to new settings by simulating the same process within the development data. For the Cox models, the baseline survival function was re-estimated in the period 1 data, then used to apply the models to the held-out period 2 data.

An assessment of performance heterogeneity can also be embedded within split-sample approaches in very large datasets[58], but IECV may have advantages in estimating sub-group performance metrics due to the higher number of held-out predictions it generates and the precision this provides.

For IECV, women entering the cohort during the first decade had their follow-up time truncated at 31 December 2009 and their status updated accordingly to preserve the temporal split. After IECV and using the individual-level predictions generated

therefrom, region-level estimates of performance metrics were calculated and then pooled using random-effects meta-analysis[57,63,64] with the Hartung-Knapp-Sidik-Jonkmann method[65] to provide a pooled meta-estimate, a 95% confidence interval, and a 95% prediction interval. The latter provides a likely range of model performance should the model be applied to a different population[57].

The IECV process estimates transportability, but not overfitting. Overfitting describes the tendency of a model to fit the development data well but perform poorly on a validation dataset – this is largely a function of dataset size (e.g. smaller datasets have higher risk of overfitting) but can also be influenced by the flexibility of the modelling technique used. Reducing overfitting can be attempted during model estimation, such as with regularisation, or post-modelling. For the latter, the approach of van Houwelingen and Le Cessie can be used as part of internal validation, or a bootstrap approach based on estimating and re-estimating the calibration slope. Given the size of the datasets available in comparison to the minimum sample size requirements for each model, overfitting was deemed not a significant concern for the modelling in QResearch data (see below section).

Model fitting – Cox proportional hazards and competing risks regression

Non-linear relationships between continuous variables and outcomes can be handled using restricted cubic splines or fractional polynomials[66]. Due to their relative ease of expression in a model equation and subsequent implementation, fractional polynomials were selected[67]. Compared to regular polynomial equations, fractional polynomials can include logarithms and non-integers as powers. Fractional polynomial terms for age, body mass index, and Townsend deprivation score were identified for each endpoint using the complete case data in the interests of computational burden and were used throughout all

modelling steps. This process was performed separately for Cox and competing risks models to permit different predictor-outcome associations.

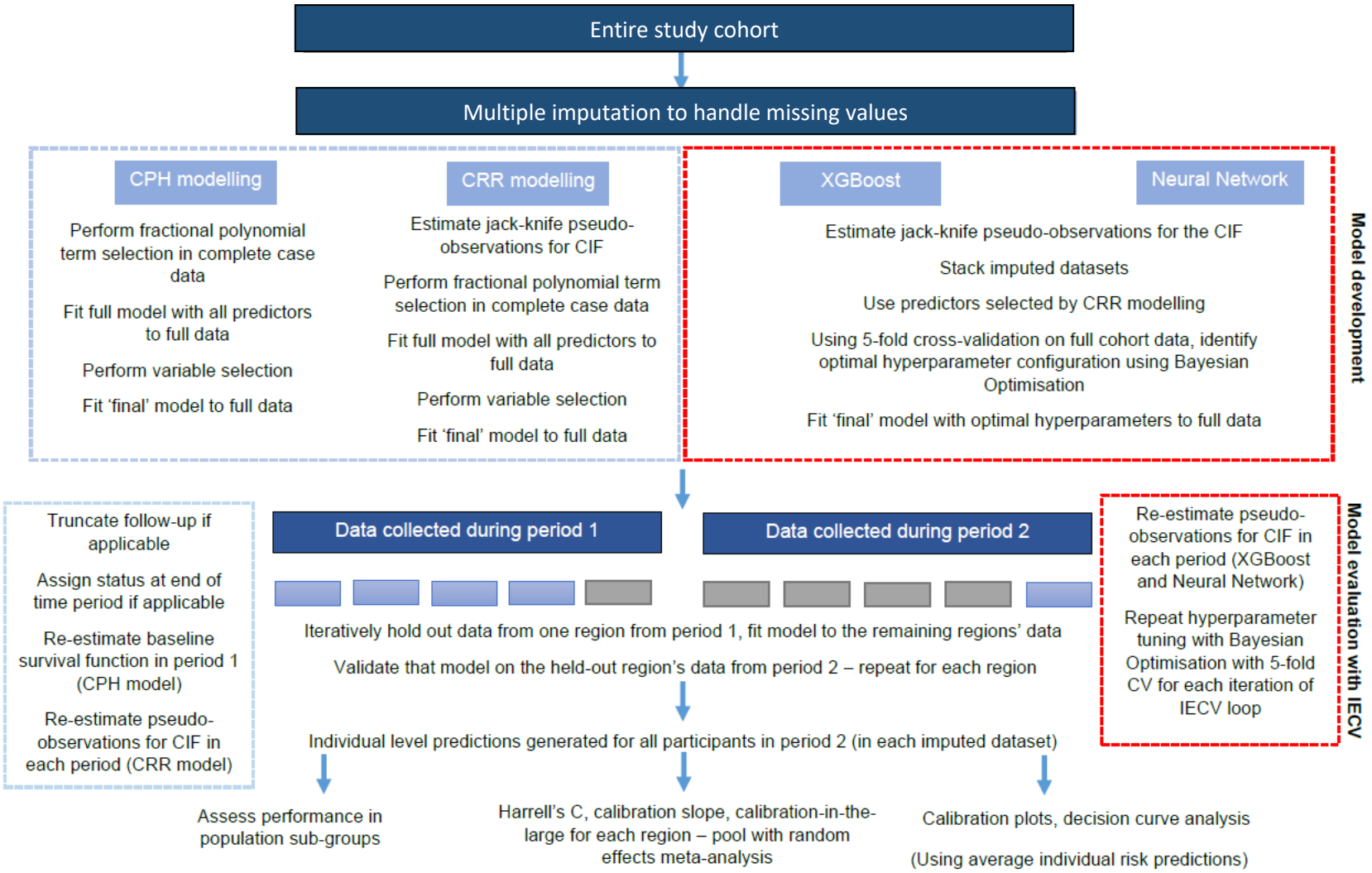

Figure 2.1. Internal-external cross-validation (IECV) strategy used to assess the performance of each model developed. CIF = cumulative incidence function, CRR = competing risks regression. Rather than the CIF, alternatively the Kaplan-Meier failure-function can be targeted with the pseudo-observations approach for a cause-specific framework.

Full models using all candidate predictors and pre-specified interaction terms were initially fitted to the entire cohort – this comprised fitting the model to each imputed dataset and combination of point estimates and their standard errors in accordance with Rubin's rules[27,68]. Baseline survival functions at 10 years were estimated in each imputed dataset with continuous variables set to their means, and categorical variables set to zero, then pooled according to Rubin's rules[27]. Rubin's rules provide a simple approach to pool results across imputations – for point estimates this is simply an average but pooling of standard errors considers the variance within and between results from each imputation.

Selecting predictors for inclusion in a final model is an additional consideration. Several approaches to predictor selection from a pool of candidate variables are available – these can be motivated by a desire to remove 'non-informative' predictors, retain only predictors that are deemed most useful at a pre-defined level of significance, or attain a more parsimonious (easier to implement) model. These include data-dependent 'automatic' selection methods such as the least absolute shrinkage operator (LASSO) which can also perform coefficient shrinkage, or methods such as stepwise selection or 'best subset' selection.

In this thesis, a non-automated predictor selection approach was used for each regression model. Initially, a 'full model' with all candidate predictors (including fractional polynomials and pre-specified interactions) was fitted. Continuous variables/fractional polynomials were retained if associated with $p<0.01$, as were binary variables associated with exponentiated coefficients <0.9 or >1.1 and p-value<0.01. Binary categorical variables were retained if the non-reference category was associated an exponentiated coefficient <0.9 or >1.1 at $p<0.01$; multilevel categorical variables required at least two levels to be associated with the same. Thereafter, the 'final model' was fitted to the cohort. This approach starts with a maximally complex model, and then considers the statistical

and clinical significance of variables (magnitude and strength of association with the outcome) to derive a model.

Model development – neural networks and XGBoost

Categorical variables were converted to dummies for both machine learning approaches. Values of continuous variables were min-max scaled (restrained between 0 and 1) for the neural network as the fitting dynamics can be affected by variable scaling but left unscaled for XGBoost. For the purposes of benchmarking, and lack of clear guidance on optimal predictor selection approaches for the machine learning models explored, the same predictors selected for the competing risks regression models were used for the machine learning models for each endpoint, unless the Cox proportional hazards model was clearly superior on decision curve analysis[69].

Both of the selected machine learning methods have hyperparameters which affect the fitting process but cannot be directly estimated from the dataset itself, e.g. the number of hidden layers in a neural network, or the number of boosting rounds in XGBoost. Hyperparameter tuning refers to the identification of optimal architectures and settings. The classic approach of grid search involves the exhaustive trial of all possible combinations of hyperparameters using pre-specified values. However, this may be computationally intensive as the number of potential combinations may be vast and may be subject to error or bias in human-selected specific values. Random grid search may be more efficient by stochastically sampling over the hyperparameter space, but again may be intensive if each model takes long to train.

Bayesian optimisation is a sequential, model-based approach to tuning where an unknown, non-expressible objective function with unknown derivatives (here, the

relationships between hyperparameters and model performance) is approximated with a surrogate function[70]. Initial evaluations are performed, such as trialling randomly selected hyperparameter combinations and measuring the performance of the resulting model, and these are used to construct a prior over possible surrogate functions that capture the relationships between hyperparameters and model performance (often using Gaussian processes). Evaluating the posterior of this Gaussian process can provide a range of sampling points for the most promising hyperparameter value combinations, with the precise points directed by an acquisition function[70]. This is then repeated for a set number of iterations, with the Gaussian process updated with each evaluation to narrow down the hyperparameter search space towards a global optimum. For all machine learning models, the 'expected improvement' acquisition function was used.

The final consideration for the machine learning models was missing data handling. Unlike the regression models, these do not have linear sets of coefficients with standard errors that can be combined using Rubin's rules. To permit full use of imputed data by all modelling methods, stacked imputations were used to fit the machine learning models. Here, the 5 (or 50, for endpoint 3) imputations were stacked to form a single 'long' dataset for development. In IECV, the performance of the machine learning models was estimated in each (period 2-derived) imputed dataset separately, and metrics combined in accordance with Rubin's rules.

Performance metrics and performance heterogeneity

Model performance was assessed in terms of discrimination, calibration, and clinical utility[14]. Discrimination refers to the ability of a model to distinguish between individuals that experience the outcome and do not[14]. Calibration refers to the agreement between a

model's predicted probabilities and those observed[71]. Clinical utility refers to the usefulness of a model to inform decision making[69].

Using the individual-level predictions generated from IECV, region-level estimates for Harrell's C-index, calibration slope, calibration-in-the-large were estimated for all models. Royston & Sauerbrei's D statistic and R^2_D [72,73] were estimated for the Cox proportional hazards models in each scenario (these are not estimable for the other methods used). For competing risks models (regression or machine learning), inverse probability of censoring weighting was applied during estimation of Harrell's C-index[74].

Smoothed calibration plots were generated using the individuals' mean predicted risk across all imputed datasets – these were plotted with a running smoother against jack-knife pseudo-observations for the Kaplan-Meier failure function or the Aalen-Johansen CIF as appropriate to visualise alignment between predicted and observed probabilities across the risk spectrum[75,76]. Mean predicted risks for each individual were also used for decision curve analysis – this compared the clinical utility of all models to default 'treat all' and 'treat none' scenarios and accounted for competing risks. More commonly seen in the literature regarding time-to-event modelling are calibration plots in which individuals are grouped into risk groups (e.g. by twentieth of predicted risk) but the number of groups is arbitrary and individuals within each group can be heterogeneous[66]. These were generated and shown at points in some Results chapters for points of interest rather than as the primary qualitative calibration measures.

Summary discrimination or calibration metrics alone do not show whether a model leads to better decision making. Decision curve analysis can assess clinical utility of using models to inform treatment decisions in terms of plotting the net benefit against risk thresholds[69]. Net benefit is a weighted measure of the potential benefits ('true positives')

and harms ('false positives') of using a model to make decisions across a range of threshold probabilities:

$$Net\ benefit = \frac{True\ positives}{N} - \frac{False\ positives}{N}\ \text{x}\ \left(\frac{Pt}{1-Pt}\right)$$

Pt = threshold probability at which a patient is deemed 'positive', N = number of individuals[69]

Last, understanding heterogeneity in model performance (e.g. in different ethnic groups) is often neglected but important to consider with regards to potential 'algorithmic bias'[9,77], and the potential for differential factor-outcome associations in different sub-groups. As well as regional and overall meta-estimates from IECV, Harrell's C, calibration slope and calibration-in-the-large were estimated by ethnic group and age group. Decision curve analysis was repeated in different sub-groups depending on the endpoint (justified and described in relevant Results chapters, **3-5**).

Random effects meta-regression[57] was also used for each model to quantify the amount of variation in regional summary performance measures that were attributable to inter-region heterogeneity in terms of age profiles, body mass index, deprivation (all the standard deviation of the relevant measurements), and ethnic diversity (defined simply as the percentage of non-white individuals). The I^2 estimated the overall amount of variability in model performance due to inter-regional heterogeneity, and the R^2 estimated the amount attributable to the aforementioned variables.

Minimum sample size calculations

These used the methods of Riley, et al.[78] for each endpoint. The assumptions used incidence/mortality data from Cancer Research UK[61], and mean follow-up of 6 years as reported in a population-based cohort study of men aged 40-75 years conducted prior to this thesis using data from QResearch[81]. For each calculation, 15% of the maximum permitted Cox-Snell R^2 was used, as was a total of 100 candidate predictors (to permit fractional polynomials and interaction terms).

For endpoint 1, with a Cox-Snell R^2 of 0.072 and an age-standardised annual breast cancer diagnosis rate of 0.01665 (166.5/100,000), the minimum sample size was estimated as 11,994 individuals with 1,199 outcome events (11.98 events per predictor parameter).

For endpoint 2, with a Cox-Snell R^2 of 0.0045 and an age-standardised annual breast cancer mortality rate of 0.000334 (33.4/100 000), the minimum sample size was estimated as 199,500 individuals with 400 outcome events (4 events per predictor parameter).

For endpoint 3, with a Cox-Snell R^2 of 0.085 and an annual mortality rate of 0.024 (estimated from a 76% 10-year survival rate), the minimum sample size was estimated as 10,080 individuals with 1,452 outcome events (14.52 events per predictor parameter).

There are no established methods to estimate minimum sample size for the machine learning models explored. There is some evidence from simulation studies that they may have up to 10-times the data requirement, but this is limited to binary outcome models[79].

Statistical software and code

Data processing, imputation, regression modelling, and all model evaluation used Stata V17. Machine learning model fitting and optimisation used R. The analysis code for all three endpoints is available via the following GitHub repositories:

https://github.com/AshDF91/OX129_incident_breast_cancer_modelling

https://github.com/AshDF91/Breast-cancer-mortality-prediction-popn-level-

https://github.com/AshDF91/Breast-cancer-prognosis

Patient and public involvement and engagement (PPIE)

At the start of the project, two PPIE exercises were used. First, two women with a personal history of breast cancer were consulted regarding the project's scope, candidate predictors and perceived usefulness to relevant stakeholders. They provided feedback on these and co-created a project lay summary. Second, the modelling parts of this thesis were presented at an Oxfordshire breast cancer support group to gain feedback on the project's aims, conduct, and elicit thoughts regarding the potential benefits and harms of using algorithmic approaches to inform breast cancer screening, prevention and care.

Study review and approval

These modelling studies using the QResearch database were reviewed and approved by the QResearch scientific committee (reference: OX129). The ethical approval for the QResearch database is from the Derby Research Ethics Committee (reference: 18/EM/0400).

Chapter references

1 van Royen FS, Moons KGM, Geersing G-J, et al. Developing, validating, updating and judging the impact of prognostic models for respiratory diseases. *Eur Respir J* 2022; : 2200250.

2 Kleinrouweler CE, Cheong-See FM, Collins GS, et al. Prognostic models in obstetrics: available, but far from applicable. *Am J Obstet Gynecol* 2016; **214**: 79-90.e36.

3 Lecun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; **521**: 436–44.

4 Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *N Engl J Med* 2019; **380**: 1347–58.

5 Dhiman P, Ma J, Navarro CA, et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. *J Clin Epidemiol* 2021; **138**: 60–72.

6 Dhiman P, Ma J, Andaur Navarro CL, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. *BMC Med Res Methodol* 2022; **22**. DOI:10.1186/s12874-022-01577-x.

7 Andaur Navarro CL, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ* 2021; **375**. DOI:10.1136/BMJ.N2281.

8 Kundu S. AI in medicine must be explainable. *Nat Med* 2021; **27**: 1328.

9 Li J, Bzdok D, Chen J, et al. Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Sci Adv* 2022; **8**:

1812.

10 Li Y, Sperrin M, Ashcroft DM, et al. Consistency of variety of machine learning and statistical models in predicting clinical risks of individual patients: longitudinal cohort study using cardiovascular disease as exemplar. *BMJ* 2020; **371**. DOI:10.1136/BMJ.M3919.

11 Adhikari S, Normand SL, Bloom J, et al. Revisiting performance metrics for prediction with rare outcomes. *Stat Methods Med Res* 2021; **30**: 2352–66.

12 Smith H, Sweeting M, Morris T, et al. A scoping methodological review of simulation studies comparing statistical and machine learning approaches to risk prediction for time-to-event data. *Diagnostic Progn Res* 2022; **6**. DOI:10.1186/S41512-022-00124-Y.

13 Alaa AM, Gurdasani D, Harris AL, et al. Machine learning to guide the use of adjuvant therapies for breast cancer. DOI:10.1038/s42256-021-00353-8.

14 Van Calster B, Wynants L, Timmerman D, et al. Predictive analytics in health care: how can we know it works? *J Am Med Inform Assoc* 2019; **26**: 1651–4.

15 Mannu GS, Wang Z, Broggio J, et al. Invasive breast cancer and breast cancer mortality after ductal carcinoma in situ in women attending for breast screening in England, 1988-2014: population based observational cohort study. *BMJ* 2020; **369**. DOI:10.1136/BMJ.M1570.

16 Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. *Stat Med* 2004; **23**: 1111–30.

17 Louro J, Posso M, Hilton Boon M, et al. A systematic review and quality assessment of individualised breast cancer risk prediction models. *Br J Cancer*

2019; **121**: 76–85.

18 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**. DOI:10.1136/bmjopen-2015-007825.

19 Escala-Garcia M, Morra A, Canisius S, et al. Breast cancer risk factors and their effects on survival: a Mendelian randomisation study. *BMC Med* 2020; **18**. DOI:10.1186/S12916-020-01797-2.

20 Karapanagiotis S, Pharoah PDP, Jackson CH, et al. Development and External Validation of Prediction Models for 10-Year Survival of Invasive Breast Cancer. Comparison with PREDICT and CancerMath. *Clin Cancer Res* 2018; **24**: 2110–5.

21 Han H, Guo W, Shi W, et al. Hypertension and breast cancer risk: a systematic review and meta-analysis. *Sci Rep* 2017; **7**. DOI:10.1038/SREP44877.

22 Liu SS, Ma XF, Zhao J, et al. Association between nonalcoholic fatty liver disease and extrahepatic cancers: a systematic review and meta-analysis. *Lipids Health Dis* 2020; **19**. DOI:10.1186/S12944-020-01288-6.

23 Harrison S, Davies AR, Dickson M, et al. The causal effects of health conditions and risk factors on social and socioeconomic outcomes: Mendelian randomization in UK Biobank. *Int J Epidemiol* 2020; **49**: 1661–81.

24 Thet Z, Lam AK, Ranganathan D, et al. Cancer risks along the disease trajectory in antineutrophil cytoplasmic antibody associated vasculitis. *Clin Rheumatol* 2020; **39**: 2501–13.

25 Jia Y, Li F, Liu YF, et al. Depression and cancer risk: a systematic review and

meta-analysis. *Public Health* 2017; **149**: 138–48.

26 White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med* 2011; **30**: 377–99.

27 Rubin DB. Multiple imputation for nonresponse in surverys. 1987 DOI:10.1002/9780470316696.

28 Tsvetanova A, Sperrin M, Peek N, et al. Missing data was handled inconsistently in UK prediction models: a review of method used. *J Clin Epidemiol* 2021; **140**: 149–58.

29 Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009; **338**: 157–60.

30 Hughes RA, Heron J, Sterne JAC, et al. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *Int J Epidemiol* 2019; **48**: 1294–304.

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

32 Madley-Dowd P, Hughes R, Tilling K, et al. The proportion of missing data should not be used to guide decisions on multiple imputation. *J Clin Epidemiol* 2019; **110**: 63–73.

33 (PDF) Imputation and Missing Indicators for handling missing data in the development and implementation of clinical prediction models: a simulation study. https://www.researchgate.net/publication/361559019_Imputation_and_Missing_Indicators_for_handling_missing_data_in_the_development_and_implementation

_of_clinical_prediction_models_a_simulation_study (accessed July 8, 2022).

34 Austin PC, Lee DS, Fine JP. Introduction to the Analysis of Survival Data in the Presence of Competing Risks. *Circulation* 2016; **133**: 601–9.

35 van Geloven N, Giardiello D, Bonneville EF, et al. Validation of prediction models in the presence of competing risks: a guide through modern methods. *BMJ* 2022; **377**: e069249.

36 Ramspek CL, Teece L, Snell KIE, et al. Lessons learnt when accounting for competing events in the external validation of time-to-event prognostic models. *Int J Epidemiol* 2022; **51**: 615–25.

37 Putter H, Schumacher M, van Houwelingen HC. On the relation between the cause-specific hazard and the subdistribution rate for competing risks data: The Fine-Gray model revisited. *Biom J* 2020; **62**: 790–807.

38 Austin PC, Steyerberg EW, Putter H. Fine-Gray subdistribution hazard models to simultaneously estimate the absolute risk of different event types: Cumulative total failure probability may exceed 1. *Stat Med* 2021; **40**: 4200–12.

39 Koller MT, Raatz H, Steyerberg EW, et al. Competing risks and the clinical community: irrelevance or ignorance? *Stat Med* 2012; **31**: 1089–97.

40 Lau B, Cole SR, Gange SJ. Competing risk regression models for epidemiologic data. *Am J Epidemiol* 2009; **170**: 244–56.

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

42 Andersen PK, Pohar Perme M. Pseudo-observations in survival analysis. *Stat Methods Med Res* 2010; **19**: 71–99.

43 Chen T, Guestrin C. XGBoost: A scalable tree boosting system. *Proc ACM SIGKDD Int Conf Knowl Discov Data Min* 2016; 13-17-August-2016: 785–94.

44 Chen T, He T. xgboost: eXtreme Gradient Boosting. 2022.

45 GitHub - dmlc/xgboost: Scalable, Portable and Distributed Gradient Boosting (GBDT, GBRT or GBM) Library, for Python, R, Java, Scala, C++ and more. Runs on single machine, Hadoop, Spark, Dask, Flink and DataFlow. https://github.com/dmlc/xgboost (accessed Aug 1, 2022).

46 Deep Learning with R. https://www.manning.com/books/deep-learning-with-r (accessed Aug 8, 2022).

47 Yala A, Lehman C, Schuster T, et al. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology* 2019; **292**: 60–6.

48 Katzman JL, Shaham U, Cloninger A, et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol* 2018; **18**. DOI:10.1186/S12874-018-0482-1.

49 Van Der Ploeg T, Datema F, Baatenburg De Jong R, et al. Prediction of survival with alternative modeling techniques using pseudo values. *PLoS One* 2014; **9**. DOI:10.1371/JOURNAL.PONE.0100234.

50 Mohammad MA, Olesen KKW, Koul S, et al. Development and validation of an artificial neural network algorithm to predict mortality and admission to hospital for heart failure after myocardial infarction: a nationwide population-based study. *Lancet Digit Heal* 2022; **4**: e37–45.

51 Steinfeldt J, Buergel T, Loock L, et al. Neural network-based integration of polygenic and clinical information: development and validation of a prediction

model for 10-year risk of major adverse cardiac events in the UK Biobank cohort. *Lancet Digit Heal* 2022; **4**: e84–94.

52 Sarvestany SS, Kwong JC, Azhie A, et al. Development and validation of an ensemble machine learning framework for detection of all-cause advanced hepatic fibrosis: a retrospective cohort study. *Lancet Digit Heal* 2022; **4**: e188–99.

53 Weng SF, Vaz L, Qureshi N, et al. Prediction of premature all-cause mortality: A prospective general population cohort study comparing machine-learning and standard epidemiological approaches. *PLoS One* 2019; **14**: e0214365.

54 Steyerberg EW. Validation in prediction research: the waste by data splitting. *J Clin Epidemiol* 2018; **103**: 131–3.

55 Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. J. Clin. Epidemiol. 2016; **69**: 245–7.

56 Debray TPA, Vergouwe Y, Koffijberg H, et al. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2015; **68**: 279–89.

57 Riley RD, Ensor J, Snell KIE, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 2016; **353**. DOI:10.1136/BMJ.I3140.

58 Clift AK, Coupland CAC, Keogh RH, et al. Living risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study. *BMJ* 2020; **371**. DOI:10.1136/BMJ.M3731.

59 Simpson CR, Robertson C, Kerr S, et al. External validation of the QCovid risk

prediction algorithm for risk of COVID-19 hospitalisation and mortality in adults: national validation cohort study in Scotland. *Thorax* 2022; **77**: 497–504.

60 de Jong VMT, Rousset RZ, Antonio-Villa NE, et al. Clinical prediction models for mortality in patients with covid-19: external validation and individual participant data meta-analysis. *BMJ* 2022; **378**: e069881.

61 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).

62 Finlayson SG, Subbaswamy A, Singh K, et al. The Clinician and Dataset Shift in Artificial Intelligence. *N Engl J Med* 2021; **385**: 283–6.

63 Austin PC, van Klaveren D, Vergouwe Y, et al. Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. *J Clin Epidemiol* 2016; **79**: 76–85.

64 Wynants L, Riley RD, Timmerman D, et al. Random-effects meta-analysis of the clinical utility of tests and prediction models. *Stat Med* 2018; **37**: 2034–52.

65 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**. DOI:10.1186/1471-2288-14-25.

66 Harrell, FE. Regression Modeling Strategies. 2015. DOI:10.1007/978-3-319-19425-7.

67 Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. *Int J Epidemiol* 1999; **28**: 964–74.

68 Austin PC, Lee DS, Ko DT, White IR. Effect of Variable Selection Strategy on the Performance of Prognostic Models When Using Multiple Imputation. *Circ Cardiovasc Qual Outcomes* 2019; **12**. DOI:10.1161/CIRCOUTCOMES.119.005927.

69 Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016; **352**. DOI:10.1136/BMJ.I6.

70 Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. *arXiv* 2012; : arXiv:1206.2944.

71 Van Calster B, McLernon DJ, Van Smeden M, et al. Calibration: The Achilles heel of predictive analytics. *BMC Med* 2019; **17**: 230.

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

73 Royston P, Group C. Explained variation for survival models. 2006.

74 Wolbers M, Blanche P, Koller MT, et al. Concordance for prognostic models with competing risks. *Biostatistics* 2014; **15**: 526.

75 Royston P. Tools for checking calibration of a Cox model in external validation: Approach based on individual event probabilities. *Stata J* 2014; **14**: 738–55.

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

77 Gichoya JW, Banerjee I, Bhimireddy AR, et al. AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit Heal* 2022; **4**: e406–14.

78 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.

79 Van Der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol* 2014; **14**. DOI:10.1186/1471-2288-14-137.

80 Vinogradova Y, Coupland C, Hippisley-Cox J. Use of hormone replacement therapy and risk of breast cancer: nested case-control studies using the QResearch and CPRD databases. *BMJ* 2020; 371:m3873.

81 Clift AK, Coupland CAC, Hippisley-Cox J. Prostate-specific antigen testing and opportunistic prostate cancer screening: a cohort study in England, 1998-2017. *Br J Gen Pract* 2021; 71(703):e157-e165.

Chapter Three

Prediction model development and evaluation – 10-year risk of incident breast cancer diagnosis

Summary

The previous chapter detailed the framework for developing and then evaluating clinical prediction models for this thesis using QResearch and its linked databases. This chapter summarises previous work done in building models to predict incident breast cancer diagnosis, and presents the results of four modelling approaches applied to this risk trajectory.

Introduction

Risk-based breast screening is generally envisioned to be predicated on accurate risk assessment for incident breast cancer diagnosis[1,2]. Several models exist which estimate such risks, as summarised in **Chapter 1**, and indeed some are being used in prospective trials of risk-adapted screening strategies such as the use of Breast Cancer Surveillance Consortium (BCSC) model[3,4] in the adaptive WISDOM trial[5]. However, two recent

systematic reviews have demonstrated uncertainty regarding the ability of currently available models to guide screening decisions[6,7].

The review of Louro, et al.[6] appraised 24 studies published up to February 2018 using the ISPOR-AMCP-NPC questionnaire, assessing them in terms of: a) internal and external validation, b) bias due to study design affecting risk estimates, c) limitations in data inputs, d) appropriateness of model analysis, e) reporting bias, f) interpretation bias, and g) conflicts of interest. Model discrimination, summary calibration (E/O ratio) and net reclassification index results were extracted from each study. This review found it challenging to recommend any model to guide population-level risk-based screening, mostly due to uncertainty on the adequacy of model discrimination, risk of bias in data inputs (e.g. data from self-report questionnaires), and difficulty of comparing summary metrics across heterogeneous populations.

The more recent systematic review of Zheng, et al.[7] appraised 47 clinical prediction models from 40 development studies published up to December 2021 using the prediction model risk of bias assessment tool (PROBAST)[8]. This review found that all previously developed incident breast cancer models identified were at high risk of bias and could not be recommended for use in routine screening programmes[7]. This was primarily due to factors such as inappropriate model performance evaluation, handling of continuous predictors with dichotomisation, missing data handling, and the lack of assessment of optimism in model performance. This review also highlighted that a minority of models underwent internal validation (n=9), and that model development and evaluation was inappropriately heavily skewed towards Caucasian women[7].

Limitations of these systematic reviews include the omission of some studies such as QCancer-10-year (Breast)[9], non-coverage of genetic prediction models, and lack of

consideration of clinical utility (driven by lack of reporting in the primary studies). As discussed in the final section of **Chapter 2**, whilst it is uncertain if stratifying women for risk-based screening in terms of incident diagnosis risk is the best way to inform mortality-reducing interventions, currently available literature demonstrates the need for robust development and evaluation of any models seeking to do so[6,7,10,11].

Most if not all currently available incident breast cancer prediction models are at high risk of bias, rendering their applicability to population risk-based screening uncertain and potentially harmful. Cognisant of best practice guidance[12,13] and the limitations of previous work, this chapter uses large-scale, population-representative linked electronic healthcare datasets to develop clinical prediction models for incident breast cancer. The evaluation strategy seeks to robustly assess performance, heterogeneity, and stress-test the models to identify aspects of poorer performance.

Study population

As per **Chapter 2**, the study population was women aged 20-90 years of age at entry into the QResearch database, without previous recorded diagnosis of breast cancer or ductal carcinoma *in situ*, contributing follow-up between 1st January 2000 and 31st December 2020. After excluding women with breast cancer (GP/HES datasets, n=152,870) or ductal carcinoma *in situ* (n=5,409) diagnoses recorded prior to cohort entry, the final study cohort comprised the same 11,626,969 women.

The incremental yield of using linked data sources is summarised in terms of crude incidence rates in **Table 3.1**, and the final cohort is summarised in **Table 3.2**. During a total of 69,310,127 person-years of follow-up, there were 142,712 incident breast cancer

diagnosis (1.23%), and 684,580 deaths from non-breast cancer-related causes (competing events, 5.89%). The median follow-up was 3.74 years (interquartile rage 1.60 to 8.44 years), and 21.36% of women had 10 years or more follow-up. The crude incidence rate overall was 20.59 per 10,000 person-years (95% CI: 20.48 to 20.70). By comparing age group-specific breast cancer incidence rates to national incidence statistics, it appeared that overall, the QResearch cohort is broadly representative of UK population in terms of breast cancer risk (**Table 3.3**).

From the perspective of the temporally distinct sub-cohorts for IECV: in the Period 1 sub-cohort, there were 59,826 incident breast cancers (crude incidence rate 20.07, 95% CI: 19.86 to 20.18). In the Period 2 sub-cohort, there were 26,538 incident breast cancers (crude incidence rate 15.22, 95% CI: 15.04 to 15.40). Regional and ethnic group-specific incidence rates are summarised in **Tables 3.4 & 3.5**, respectively.

Model development

Non-linear fractional polynomials were selected for age and BMI for the Cox proportional hazards model. For the competing risks regression model, age and Townsend deprivation score had non-linearities selected (**Figure 3.1**). The final Cox and competing risks regression models are displayed as their exponentiated coefficients in **Figures 3.2** and **3.3**, respectively, and in full in **Tables 3.6** and **3.7**, respectively. Final architectures for the XGBoost and neural network models are summarised in **Table 3.8**. Interaction terms were considered between age and family history of breast cancer, and age and BMI.

Data source(s)	Number of incident breast cancers	Person-years	Crude incidence rate per 10,000 (95% CI)
GP data only	118,598	69,407,582.6	17.09 (16.99 to 17.18)
GP + Cancer Registry	137,623	69,323,185.1	19.85 (19.75 to 19.96)
GP + Cancer Registry + HES	142,027	69,310,127.4	20.49 (20.39 to 20.60)
GP + Cancer Registry + HES + Death Register	142,712	69,310,127.4	20.59 (20.48 to 20.70)

Table 3.1. Crude incidence rates for the outcome of interest (incident breast cancer diagnosis) in the study cohort, based on case ascertainment from the different linked data assets. GP = general practice, HES = hospital episodes statistics. Follow-up was from date of cohort entry until date of breast cancer diagnosis (earliest on the relevant data assets), or censoring (left practice, reached study end alive, died from other, non-breast cancer-related cause).

Parameter	Category	Overall study cohort (Column %)	Period 1 sub-cohort (1st Jan 2000-31st Dec 2009) (Column %)	Period 2 sub-cohort (1st Jan 2010 – 31st Dec 2020) (Column %)
Total individuals		11,626,969	6,151,399	5,475,570
Age at entry	Mean (SD) Median (IQR)	41.78 (18.13) 36 (27–53)	42.95 (18.44) 38 (28-55)	40.48 (17.67) 35 (26-51)
Body mass index at entry	Mean (SD) Median (IQR)	25.37 (5.46) 24.2 (21.5–28.1)	25.02 (5.09) 24 (21.4-27.5)	25.68 (5.74) 24.4 (21.5-28.7)
Townsend deprivation score	Mean (SD)	0.71 (3.23)	0.50 (3.24)	0.94 (3.21)
Ethnic group*	White	6,168,419 (53.05)	2,751,371 (44.73)	3,417,048 (62.41)
(self-reported)	Indian	263,564 (2.27)	102,623 (1.67)	160,941 (2.94)
	Pakistani	145,168 (1.25)	55,542 (0.90)	89,626 (1.64)
	Bangladeshi	99,097 (0.85)	41,738 (0.68)	57,359 (1.05)
	Other Asian	189,635 (1.63)	66,538 (1.08)	123,097 (2.25)
	Black Caribbean	117,223 (1.01)	59,953 (0.97)	57,270 (1.05)
	Black African/Other Black	328,497 (2.83)	130,135 (2.12)	198,362 (3.62)
	Chinese	132,583 (2.82)	39,643 (0.64)	92,940 (1.70)
	Other (including mixed race, Arab)	328,396 (2.82)	101,147 (1.64)	227,249 (4.15)
	Not recorded	3,854,387 (33.15)	2,802,709 (45.56)	1,051,678 (19.21)
Smoking status	Non-smoker	5,045,101 (43.39)	2,207,095 (35.88)	2,838,006 (51.83)
	Ex-smoker	1,445,584 (12.43)	655,998 (10.66)	789,586 (14.42)
	Light smoker (1-9/day)	1,318,132 (11.34)	688,109 (11.19)	630,023 (11.51)
	Moderate smoker (10-19/day)	308,372 (2.65)	173,942 (2.83)	134,430 (2.46)
	Heavy smoker (20+/day)	133,108 (1.14)	86,337 (1.40)	46,771 (0.85)
	Not recorded	3,376,672 (29.04)	2,339,918 (38.04)	1,036,754 (18.93)
Alcohol intake	Non-drinker	4,120,142 (35.44)	1,994,174 (32.42)	2,125,968 (38.83)
	Trivial <1u/day	1,580,548 (13.59)	770,767 (12.53)	809,781 (14.79)
	Light 1-2u/day	601,071 (5.17)	289,449 (4.71)	311,622 (5.69)
	Moderate 3-6u/day	246,366 (2.12)	128,708 (2.09)	117,658 (2.15)
	Heavy 7-9u/day	9,823 (0.08)	3,865 (0.06)	5,958 (0.11)
	Very Heavy >9u/day	17,467 (0.15)	2,510 (0.04)	14,957 (0.27)
	Not recorded	5,051,552 (43.45)	2,961,926 (48.15)	2,089,626 (38.16)
Benign breast disease		282,663 (2.43)	142,108 (2.31)	140,555 (2.57)
Endometriosis		151,158 (1.30)	59,717 (0.97)	91,441 (1.67)
Polycystic ovary syndrome		197,886 (1.70)	57,951 (0.94)	139,935 (2.56)
Hysterectomy		99,439 (0.86)	31,051 (0.50)	68,388 (1.25)
Previous gynaecological cancer		26,626 (0.23)	11,264 (0.18)	15,362 (0.28)

Oral contraceptive pill use (ever)		1,372,633 (11.81)	519,506 (8.45)	853,127 (15.58)
Recent oestrogen-only hormone replacement therapy	None (reference)	11,467,510 (98.63)	6,037,813 (98.15)	5,429,697 (99.16)
(<5 years since last prescription)	<1 year duration	58,156 (0.50)	40,563 (0.66)	17,593 (0.32)
	1 - 2.9 years duration	34,566 (0.30)	27,071 (0.44)	7,495 (0.14)
	3 – 4.9 years duration	25,760 (0.22)	21,024 (0.34)	4,736 (0.09)
	5 – 9.9 years duration	30,254 (0.26)	22,943 (0.37)	7,311 (0.13)
	10+ years duration	10,723 (0.09)	1,985 (0.03)	8,738 (0.16)
Past oestrogen-only hormone replacement therapy	None (reference)	11,551,573 (99.35)	6,134,698 (99.73)	5,416,875 (98.93)
(5+ years since last prescription)	<1 year duration	35,352 (0.30)	11,035 (0.18)	24,317 (0.44)
	1 - 2.9 years duration	12,685 (0.11)	3,297 (0.05)	9,388 (0.17)
	3 – 4.9 years duration	8,732 (0.08)	1,318 (0.02)	7,414 (0.14)
	5 – 9.9 years duration	13,071 (0.11)	956 (0.02)	12,115 (0.22)
	10+ years duration	5,556 (0.05)	95 (0.00)	5,461 (0.10)
Recent combined hormone replacement therapy	None (reference)	11,339,053 (97.52)	5,929,384 (96.39)	5,409,669 (98.80)
(<5 years since last prescription)	<1 year duration	85,664 (0.74)	65,937 (1.07)	19,727 (0.36)
	1 - 2.9 years duration	68,532 (0.59)	56,284 (0.91)	12,248 (0.22)
	3 – 4.9 years duration	52,565 (0.45)	44,811 (0.73)	7,754 (0.14)
	5 – 9.9 years duration	63,127 (0.54)	50,867 (0.83)	12,260 (0.22)
	10+ years duration	18,028 (0.16)	4,116 (0.07)	13,912 (0.25)
Past combined hormone replacement therapy	None (reference)	11,489,012 (98.81)	6,124,006 (99.55)	5,365,006 (97.98)
(5+ years since last prescription)	<1 year duration	48,225 (0.41)	15,520 (0.25)	32,705 (0.60)
	1 - 2.9 years duration	26,529 (0.23)	6,177 (0.10)	20,352 (0.37)
	3 – 4.9 years duration	20,172 (0.17)	2,983 (0.05)	17,189 (0.31)
	5 – 9.9 years duration	31,508 (0.27)	2,372 (0.04)	29,136 (0.53)
	10+ years duration	11,523 (0.10)	341 (0.01)	11,182 (0.20)
Family history of breast cancer		177,368 (1.53)	64,718 (1.05)	112,650 (2.06)
Family history of gynaecological cancer		36,932 (0.32)	12,349 (0.20)	24,583 (0.45)

Previous lung cancer		9,414 (0.08)	3,529 (0.06)	5,885 (0.11)
Previous haematological cancer		31,637 (0.27)	13,050 (0.21)	18,587 (0.34)
Previous thyroid cancer		6,009 (0.05)	1,981 (0.03)	4,028 (0.07)
Type 1 diabetes mellitus		20,479 (0.18)	9,782 (0.16)	10,697 (0.20)
Type 2 diabetes mellitus		311,725 (2.68)	129,256 (2.10)	182,469 (3.33)
Chronic kidney disease	None/stage 2	11,471,953 (98.67)	6,129,623 (99.65)	5,342,330 (97.57)
	Stage 3	136,585 (1.17)	15,077 (0.25)	121,508 (2.22)
	Stage 4	7,972 (0.07)	1,113 (0.02)	6,859 (0.13)
	Stage 5 (inc. end-stage renal failure/dialysis)	10,459 (0.09)	5,586 (0.09)	4,873 (0.09)
Hypertension		1,073,831 (9.24)	524,255 (8.52)	549,576 (10.04)
Ischaemic heart disease		255,299 (2.20)	147,567 (2.40)	107,732 (1.97)
Chronic liver disease		30,550 (0.26)	10,076 (0.16)	20,474 (0.37)
Systemic lupus erythematosus		14,541 (0.13)	6,483 (0.11)	8,058 (0.15)
Vasculitis		63,329 (0.54)	28,416 (0.46)	34,913 (0.64)
Psychotic condition		87,334 (0.75)	38,072 (0.62)	49,262 (0.90)
Anti-psychotic medication use (ever)		119,285 (1.03)	54,964 (0.89)	64,321 (1.17)
Thiazide use (ever)		433,488 (3.73)	233,350 (3.79)	200,138 (3.66)
Beta-blocker use (ever)		547,073 (4.71)	272,165 (4.42)	274,908 (5.02)
Renin-angiotensin-aldosterone axis inhibitor use (ever)		522,320 (4.49)	195,058 (3.17)	327,262 (5.98)
Angiotensin converting enzyme inhibitor use (ever)		441,264 (3.80)	174,109 (2.83)	267,155 (4.88)
Calcium channel blocker use (ever)		392,027 (3.37)	141,830 (2.31)	250,197 (4.57)
Tricyclic antidepressant use (ever)		520,745 (4.48)	242,492 (3.94)	278,253 (5.08)
Monoamine oxidase inhibitor use (ever)		3,328 (0.03)	2,277 (0.04)	1,051 (0.02)
Selective serotonin reuptake inhibitor use (ever)		886,309 (7.62)	263,703 (4.29)	622,606 (11.37)

Table 3.2. Summary characteristics of the study cohort overall, and in the temporally distinct sub-cohorts used for the internal-external validation. SD = standard deviation, IQR = interquartile

range, u = alcohol units. Ever-use of medications is defined as the receipt of 3 or more prescriptions in the primary care records. Ethnic group was not included as a candidate predictor for the models, but was of interest to assess model performance heterogeneity. Screening history (e.g. previous mammography) was considered as a candidate predictor, but this was excluded due to poor data quality manifesting as lower-than-expected rates of screening.

	Crude incidence rate per 10,000 person-years (95% CIs available for QResearch)	
Age group	QResearch cohort	Cancer Research UK statistics
20-24 years	0.14 (0.11 to 0.19)	0.2
25-29 years	0.82 (0.75 to 0.89)	1.1
30-34 years	2.41 (2.30 to 2.53)	3.1
35-39 years	5.87 (5.69 to 6.05)	6.6
40-44 years	11.56 (11.31 to 11.82)	12.5
45-49 years	20.07 (19.73 to 20.42)	21.5
50-54 years	29.99 (29.56 to 30.43)	28.0
55-59 years	29.59 (29.14 to 30.05)	28.5
60-64 years	36.41 (35.88 to 36.95)	33.8
65-69 years	39.71 (39.11 to 40.31)	41.2
70-74 years	35.92 (35.32 to 36.54)	37.3
75-79 years	37.68 (37.00 to 38.37)	40.3
80-84 years	40.43 (39.64 to 41.22)	43.0
85-89 years	44.66 (43.66 to 45.68)	44.8 (85+ years)*
90-94 years	45.87 (44.31 to 47.48)	44.8 (85+ years)*
95-100 years	47.14 (43.29 to 51.34)	44.8 (85+ years) *

Table 3.3. Age group-specific incidence rates for incident breast cancer in the present study cohort from QResearch (and linked databases), and national incidence rates for the same, obtained from Cancer Research UK for the years 2017-2019 https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/incidence-invasive#heading-One (accessed 23rd Sept 2022).

* = Cancer Research UK data groups together the age-bands for 85 years and above. Age group does not refer to age at entry into the QResearch database/cohort – follow-up time for an individual woman could span different age groups.

Geographical region	Number of incident breast cancers (all 4 linked data sources)	Person-years	Crude incidence rate per 10,000 (95% CI)
Overall study period (1st Jan 2000 to 31st Dec 2020)			
East Midlands	6,832	3,152,607.8	21.67 (21.16 to 22.19)
East of England	9,079	4,045,003.2	22.44 (21.99 to 22.91)
London	22,733	15,490,307	14.68 (14.49 to 14.87)
North East	5,188	2,407,428.8	21.55 (20.97 to 22.14)
North West	25,225	11,548,763	21.84 (21.57 to 22.11)
South Central	19,031	8,496,904.4	22.40 (22.08 to 22.72)
South East	14,327	5,962,283.9	24.03 (23.64 to 24.43)
South West	17,375	7,563,478.2	22.97 (22.63 to 23.32)
West Midlands	15,479	7,036,579.3	21.99 (21.65 to 22.35)
Yorkshire & Humber	7,443	3,606,770.9	20.64 (20.18 to 21.11)
Period 1 sub-cohort (1st Jan 2000 to 31st Dec 2009)			
East Midlands	3,469	1,656,397.5	20.94 (20.26 to 21.65)
East of England	4,214	1,971,056.8	21.38 (20.74 to 22.03)
London	9,353	6,253,746.4	14.96 (14.66 to 15.26)
North East	2,386	1,159,331.2	20.58 (19.77 to 21.42)
North West	9,712	4,710,039.4	20.62 (20.21 to 21.03)
South Central	7,701	3,681,012.8	20.92 (20.46 to 21.39)
South East	5,622	2,363,042.3	23.79 (23.18 to 24.42)
South West	7,160	3,265,415.4	21.93 (21.42 to 22.44)
West Midlands	6,737	3,058,172.2	22.03 (21.51 to 22.56)
Yorkshire & Humber	3,472	1,770,750.3	19.61 (18.97 to 20.27)
Period 2 sub-cohort (1st Jan 2010 to 31st Dec 2010)			
East Midlands	685	470,306.09	14.56 (13.51 to 15.70)
East of England	1,349	745,169.81	18.10 (17.16 to 19.09)

London	5,154	5,096,204.8	10.11 (9.84 to 10.39)
North East	426	310,909.42	13.70 (12.46 to 15.07)
North West	5,412	2,961,874	18.27 (17.79 to 18.77)
South Central	3,353	2,004,102.5	16.73 (16.17 to 17.31)
South East	3,241	1,726,124.6	18.78 (18.14 to 19.43)
South West	3,237	1,823,307.7	17.75 (17.15 to 18.38)
West Midlands	2,927	1,740,303.1	16.82 (16.22 to 17.44)
Yorkshire & Humber	774	572,018.22	13.53 (12.61 to 14.52)

Table 3.4. Geographical region-specific crude incidence rates for incident breast cancer in the overall study cohort, and separately for the temporally distinct Period 1 and Period 2 sub-cohorts. For the overall cohort, follow-up was from date of cohort entry until date of breast cancer diagnosis (earliest on the relevant data assets), or censoring (left practice, reached study end alive, died from other, non-breast cancer-related cause). For Period 1, follow-up was from date of cohort entry until the earliest of: date of breast cancer diagnosis (earliest on the relevant data assets), censoring (left practice, reached study end alive, died from other, non-breast cancer-related cause), or 31st December 2009. For Period 2, follow-up was from date of cohort entry until the date of breast cancer diagnosis (earliest on the relevant data assets), censoring (left practice, reached study end alive, died from other, non-breast cancer-related cause).

Ethnic group	Number of incident breast cancers (all 4 linked data sources)	Person-years	Crude incidence rate per 10,000 (95% CI)
Overall study period (1st Jan 2000 to 31st Dec 2020)			
White	92,692	42,069,859	22.03 (21.89 to 22.18)
Indian	1,672	1,441,536	11.60 (11.06 to 12.17)
Pakistani	868	866,874.24	10.01 (9.37 to 10.70)
Bangladeshi	322	583,391.35	5.52 (4.95 to 6.16)
Other Asian	1,018	896,470.19	11.36 (10.68 to 12.08)
Black Caribbean	1,321	820,045.06	16.11 (15.26 to 17.00)
Black African	1,423	1,521,252	9.35 (8.88 to 9.85)
Chinese	391	488,217.4	8.19 (7.42 to 9.08)
Other (inc. mixed race, Arab)	1,444	1,353,776.3	10.67 (10.13 to 11.23)
Period 1 sub-cohort (1st Jan 2000 to 31st Dec 2009)			
White	33,503	15,976,322	20.97 (20.75 to 21.20)
Indian	510	457,636.4	11.14 (10.22 to 12.15)
Pakistani	232	264,615.07	8.77 (7.71 to 9.97)
Bangladeshi	66	175,500.69	3.76 (2.95 to 4.79)
Other Asian	258	231,978.03	11.12 (9.84 to 12.57)
Black Caribbean	409	313,830.87	13.03 (11.83 to 14.36)
Black African	314	436,773.34	7.19 (6.44 to 8.03)
Chinese	103	112,235.54	9.18 (7.57 to 11.13)
Other (inc. mixed race, Arab)	326	334,031.07	9.76 (8.76 to 10.88)

Period 2 sub-cohort (1st Jan 2010 to 31st Dec 2010)			
White	18,404	11,248,569	16.36 (16.13 to 16.60)
Indian	441	516,519.86	8.54 (7.78 to 9.37)
Pakistani	254	306,131.19	8.30 (7.34 to 9.38)
Bangladeshi	101	204,035.66	4.95 (4.07 to 6.01)
Other Asian	346	385,084.88	8.99 (8.09 to 9.98)
Black Caribbean	337	209,323.45	16.10 (14.47 to 17.91)
Black African	567	627,519.87	9.04 (8.32 to 9.81)
Chinese	141	249,001.21	5.66 (4.80 to 6.68)
Other (inc. mixed race, Arab)	576	652,293.25	8.83 (8.14 to 9.58)

Table 3.5. Ethnic group-specific crude incidence rates for incident breast cancer in the overall study cohort, and separately for the temporally distinct Period 1 and Period 2 sub-cohorts. For the overall cohort, follow-up was from date of cohort entry until date of breast cancer diagnosis (earliest on the relevant data assets), or censoring (left practice, reached study end alive, died from other, non-breast cancer-related cause). For Period 1, follow-up was from date of cohort entry until the earliest of: date of breast cancer diagnosis (earliest on the relevant data assets), censoring (left practice, reached study end alive, died from other, non-breast cancer-related cause), or 31st December 2009. For Period 2, follow-up was from date of cohort entry until the date of breast cancer diagnosis (earliest on the relevant data assets), censoring (left practice, reached study end alive, died from other, non-breast cancer-related cause).

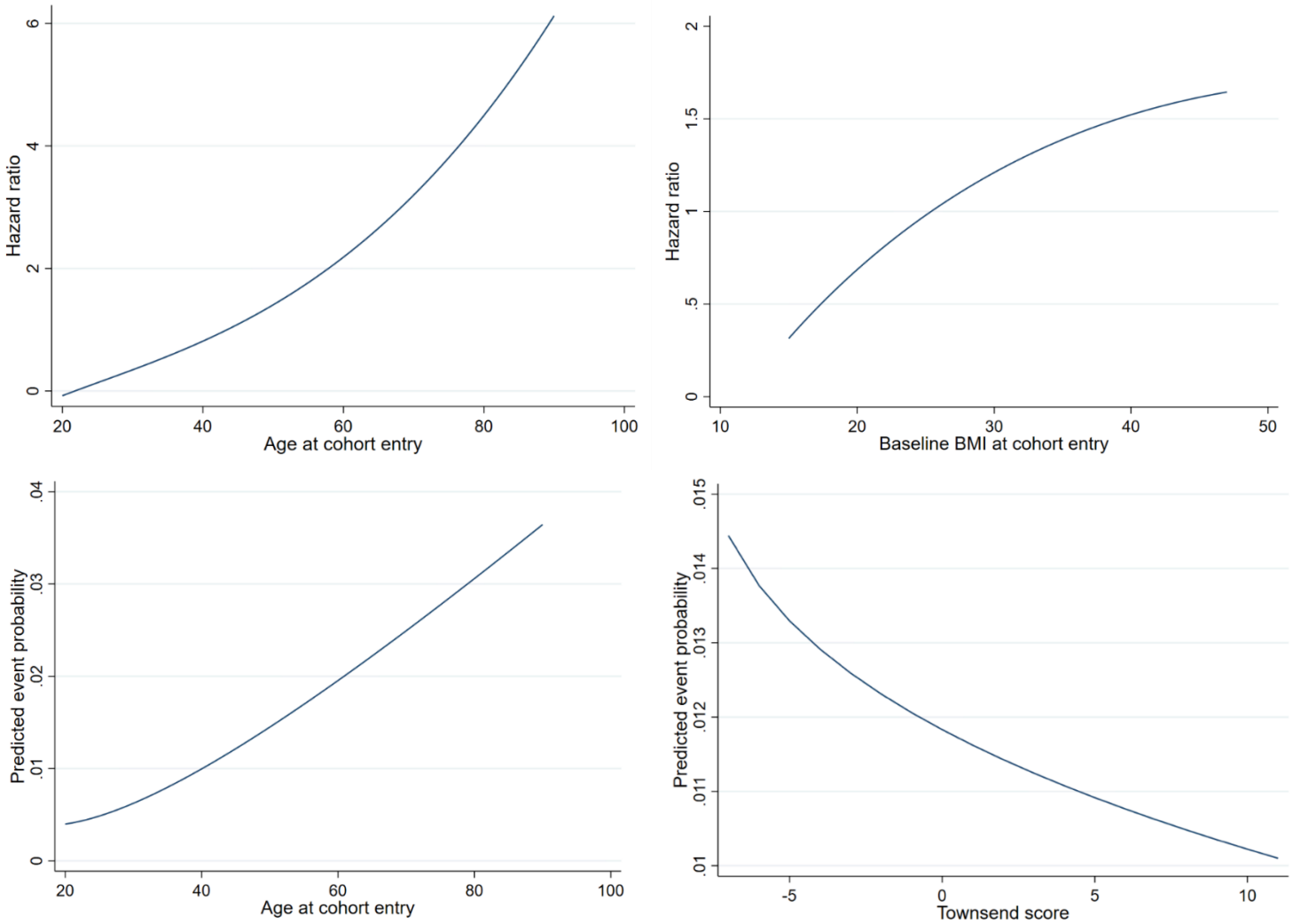

Figure 3.1. Fractional polynomial functional forms selection for age and body mass index for the final Cox proportional hazards model (top); and for age and Townsend deprivation score for the final competing risks regression model (bottom).

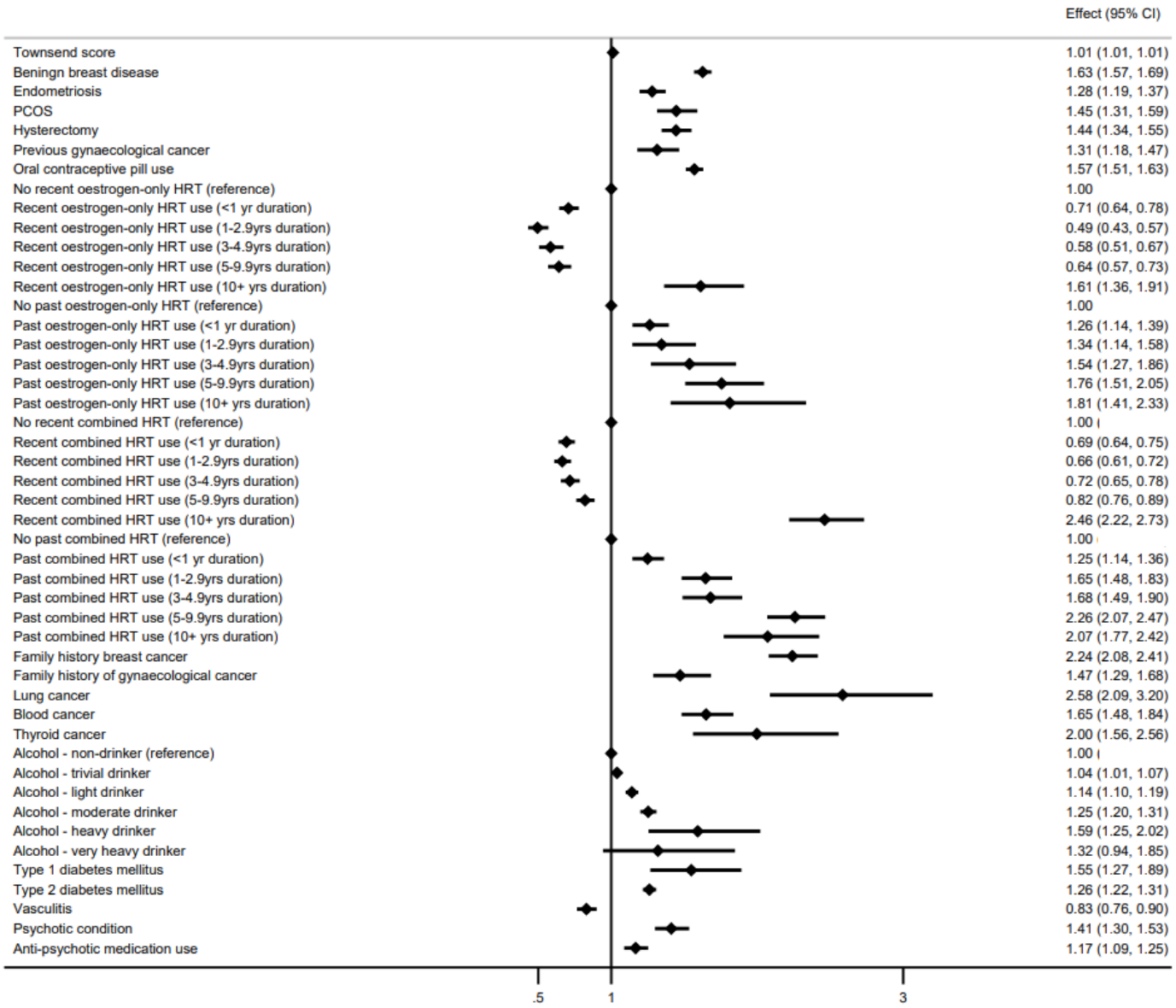

Figure 3.2. Final Cox proportional hazards model displayed as exponentiated coefficients and 95% confidence intervals. Age*family history of breast cancer interaction term not displayed due to scale. Fractional polynomial terms for age (powers -2, 3) and body mass index (-2, -2) are not plotted due to scale.

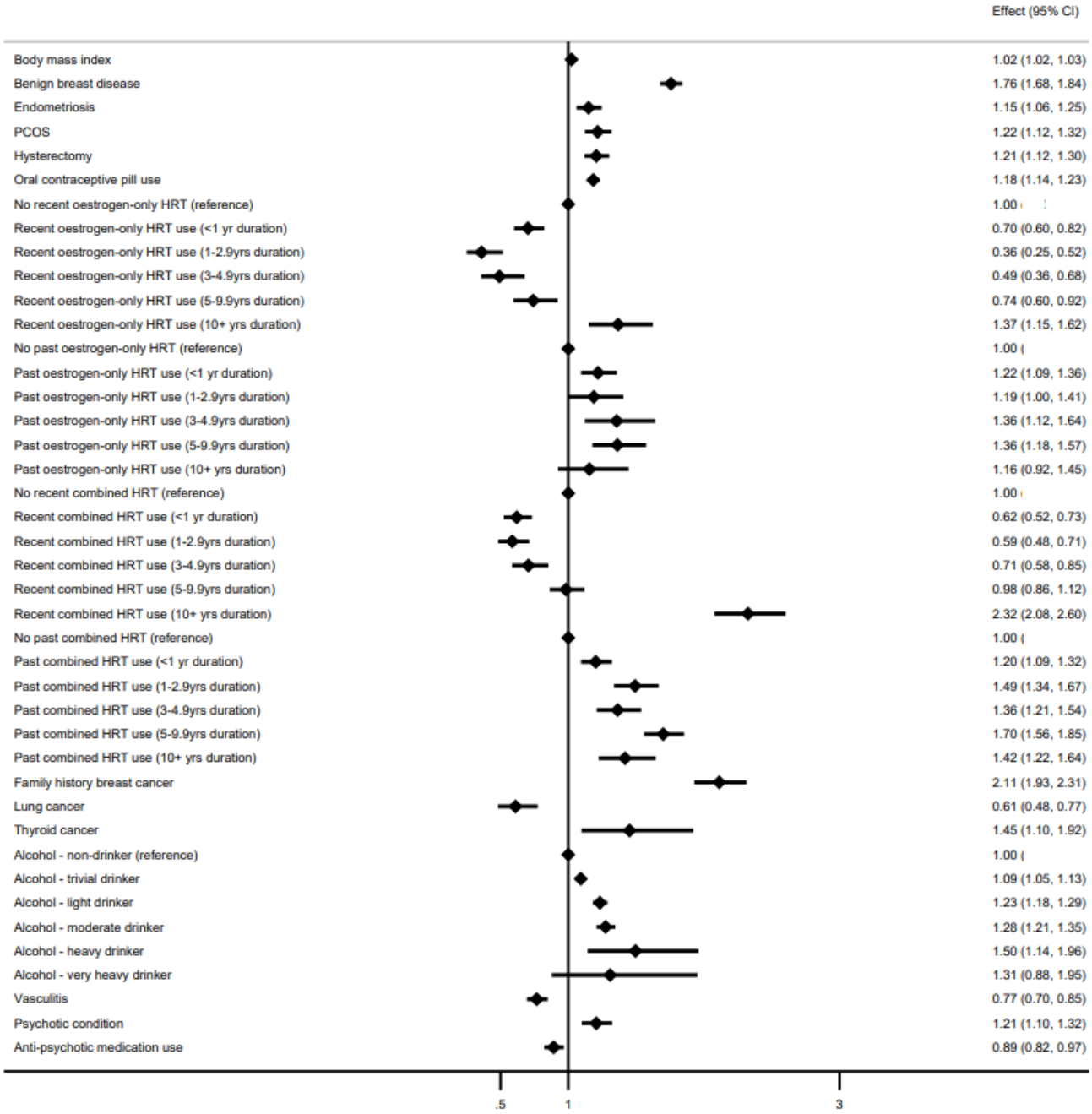

Figure 3.3. Final competing risks regression model displayed as exponentiated coefficients with 95% confidence intervals. Fractional polynomial terms for age (powers -1, -1) and Townsend deprivation score (powers 0, 0.5) are not visually displayed above due to issues of scale, nor are the interaction terms between the fractional polynomial terms for age and family history of breast cancer.

Parameter	Category/description	Coefficient
Age at entry – FP term 1	X^{-2} – 0.55476233 *X = age/10*	-21.175859
Age at entry – FP term 2	X^{3} – 76.53138187 *X = age/10*	0.00156124
Body mass index – FP term 1	X^{-2} – 0.15531752 *X = BMI/10*	0.92508765
Body mass index – FP term 2	X^{-2} * ln(X) – 0.1446226461 *X = BMI/10*	0.01100764
Benign breast disease	No (reference) Yes	0 0.4859072
Endometriosis	No (reference) Yes	0 0.24687212
Previous hysterectomy	No (reference) Yes	0 0.36714481
Previous gynaecological cancer	No (reference) Yes	0 0.27289132
Oral contraceptive pill use	No (reference) Yes	0 0.45064521
Recent oestrogen-only HRT use	None (reference) <1 year duration 1 - 2.9 years duration 3 – 4.9 years duration 5 – 9.9 years duration 10+ years duration	0 -0.34739766 -0.70404912 -0.53939226 -0.4444074 0.47773697
Past oestrogen-only HRT use	None (reference) <1 year duration 1 - 2.9 years duration 3 – 4.9 years duration 5 – 9.9 years duration 10+ years duration	0 0.23395249 0.29604515 0.42898416 0.56291697 0.59439672
Recent combined HRT use	None (reference) <1 year duration 1 - 2.9 years duration 3 – 4.9 years duration 5 – 9.9 years duration 10+ years duration	0 -0.36743462 -0.40968971 -0.33402261 -0.1980899 0.90091382
Past combined HRT use	None (reference) <1 year duration 1 - 2.9 years duration 3 – 4.9 years duration 5 – 9.9 years duration 10+ years duration	0 0.22170673 0.4983948 0.51872159 0.81466137 0.72812412
Family history of breast cancer	No (reference) Yes	0 0.80668864
Family history of gynaecological cancer	No (reference) Yes	0 0.38646923
Previous lung cancer	No (reference) Yes	0 0.94964038
Previous haematological cancer	No (reference) Yes	0 0.50046253
Previous thyroid cancer	No (reference) Yes	0 0.69174103

Alcohol intake	Non-drinker (reference)	0
	Trivial <1u/day	0.03830531
	Light 1-2u/day	0.13190495
	Moderate 3-6u/day	0.22547258
	Heavy 7-9u/day	0.46515504
	Very Heavy >9u/day	0.27743907
Type 1 diabetes mellitus	No (reference)	0
	Yes	0.43706883
Type 2 diabetes mellitus	No (reference)	0
	Yes	0.23189674
Vasculitis	No (reference)	0
	Yes	-0.18722647
History of psychotic condition	No (reference)	0
	Yes	0.34441959
Anti-psychotic medication use	No (reference)	0
	Yes	0.15437875
Interaction: Age FP term 2 * family history of breast cancer		-0.00125287
Baseline survival function at 10 years		0.9920542

Table 3.6. Final Cox proportional hazards model coefficients and baseline survival function. HRT = hormone replacement therapy, FP = fractional polynomial, BMI = body mass index (kg/m^2).

Parameter	Category/description	Coefficient
Age at entry – FP term 1	X^{-1} – 0.2355339317 *X = age/10*	-3.4088552
Age at entry– FP term 2	X^{-1} * ln(X) – 0.3405585807 *X = age/10*	-8.5843372
Body mass index		0.2376309
Townsend deprivation score – FP term 1	ln(X) + 0.1493358247 *X = (Townsend score+8)/10*	0.65996847
Townsend deprivation score – FP term 2	$X^{-0.5}$ – 0.9280516296 *X = (Townsend score+8)/10*	3.0791695
Benign breast disease	No (reference) Yes	0 0.56366448
Endometriosis	No (reference) Yes	0 0.13972739
Polycystic ovary syndrome	No (reference) Yes	0 0.19603279
Previous hysterectomy	No (reference) Yes	0 0.18804667
Oral contraceptive use	No (reference) Yes	0 0.168896
Recent oestrogen-only HRT use	None (reference) <1 year duration 1 - 2.9 years duration 3 – 4.9 years duration 5 – 9.9 years duration 10+ years duration	0 -0.35268196 -1.0312153 -0.70838053 -0.29929138 0.31261931
Past oestrogen-only HRT use	None (reference) <1 year duration 1 - 2.9 years duration 3 – 4.9 years duration 5 – 9.9 years duration 10+ years duration	0 0.19792436 0.17246012 0.30502482 0.30846467 0.14481282
Recent combined HRT use	None (reference) <1 year duration 1 - 2.9 years duration 3 – 4.9 years duration 5 – 9.9 years duration 10+ years duration	0 -0.47812067 -0.53229169 -0.34785709 -0.01700628 0.84368277
Past combined HRT use	None (reference) <1 year duration 1 - 2.9 years duration 3 – 4.9 years duration 5 – 9.9 years duration 10+ years duration	0 0.18441653 0.40065402 0.31014859 0.53037809 0.34990317
Family history of breast cancer	No (reference) Yes	0 0.74814264
Previous lung cancer	No (reference) Yes	0 -0.49262612
Previous thyroid cancer	No (reference) Yes	0 0.37289123
Alcohol intake	Non-drinker (reference) Trivial <1u/day Light 1-2u/day	0 0.08809646 0.20963305

	Moderate 3-6u/day Heavy 7-9u/day Very Heavy >9u/day	0.24302632 0.40290806 0.26883765
Vasculitis	No (reference) Yes	0 -0.26423642
History of psychotic condition	No (reference) Yes	0 0.1874511
Anti-psychotic medication use	No (reference) Yes	0 -0.11409927
Interaction: Age FP term 1 * family history of breast cancer		-8.2472493
Interaction: Age FP term 2 * family history of breast cancer		17.677752
Constant term		-5.2657479

Table 3.7. Final competing risks regression model coefficients and baseline survival function. HRT = hormone replacement therapy, u = units of alcohol, FP = fractional polynomial.

Model	Basic architecture	Hyperparameters tuned	Range explored	Final selected value
XGBoost	Tree-based booster with 'GPU_hist' method	Maximum tree depth	1 to 6	5
	Gradient-based sub-sampling	Learning rate (eta)	0 to 0.3	0.245
	'Reg:squarederror' objective	Sub-sample proportion	0.1 to 0.8	0.346
	RMSE evaluation metric	Number of boosting rounds	0 to 500	128
		Alpha (regularisation)	0 to 20	3
		Gamma (regularisation)	0 to 20	1
		Lambda (regularisation)	0 to 20	14
		Column sampling by tree	0.1 to 0.8	0.594
		Column sampling by level	0.1 to 0.8	0.763
Neural network	Feed-forward ANN with fully connected layers	Number of hidden layers	1 to 5	4
	ReLU activation functions in hidden layers	Number of nodes in each hidden layer	24 to 48	46
	Adam optimiser	Number of epochs	1 to 10	1
	Single output node with linear activation	Initial learning rate	0.001 to 0.1	0.015
	RMSE loss function			
	Batch size 8192			

Table 3.8. Hyperparameter tuning, and final configurations of the machine learning models developed to predict 10-year risk of incident breast cancer diagnosis. RMSE = root mean squared error, ANN = artificial neural network. ReLU = rectified linear unit. The final neural network had a total of 32,737 parameters. XGBoost used 50 iterations of Bayesian Optimization, the neural network used 20 iterations due to the smaller hyperparameter search space and computational expense.

Model evaluation – overall performance

The summary performance metrics estimated using IECV for all 4 models are summarised in **Table 3.9**. The competing risks regression model had the lowest discrimination (Harrell's C 0.711, 95% CI: 0.693 to 0.728, 95% PI: 0.662 to 0.760), whilst the Cox model had the highest (Harrell's C 0.782, 95% CI: 0.761 to 0.803, 95% PI: 0.712 to 0.853). The discrimination of the machine learning models was slightly lower than that for the Cox model, suggesting an overall superiority of modelling in a cause-specific framework in this specific setting.

In terms of calibration, the competing risks regression model showed marked misalignment between predicted and observed risks on both summary metrics and smoothed plots. The IECV-estimated calibration slope was 0.573 (95% CI: 0.555 to 0.592, 95% PI: 0.512 to 0.635). In contrast, the Cox model was not miscalibrated on summary measures (slope 0.962, 95% CI: 0.909 to 1.014, 95% PI: 0.797 to 0.126), and accounted for 34.9% of the variation in time to breast cancer diagnosis (95% CI: 30.9 to 38.9, 95% PI: 21.7 to 48.1). The XGBoost and neural network models showed negligible miscalibration (**Table 3.9**) regarding the same measures.

However, smoothed calibration plots showed varying patterns of (mis)calibration across the spectrum of predicted risks that were (perhaps unsurprisingly) not fully appreciable with summary metrics (**Figure 3.4**). Of note is the miscalibration of the competing risks regression model – the confidence intervals around the calibration curve at the very highest predicted risks may appear to cross implausible values; this is because pseudo-observations (used to estimate observed risks) are not constrained to be between 0 and 1[14,15].

The Cox, XGBoost and neural networks generally tended to overestimation of risk above thresholds between 6 to 10%. This was then contextualised against the predicted risk distributions (**Figure 3.5**), and corresponded to overestimation in the highest risk 1.02% to 1.75% of patients depending on the model.

Figure 3.6 demonstrates the calibration plots truncated at 6% for the purposes of interpretability, whilst the calibration of the competing risks model is shown for the entire risk spectrum. The Cox model displayed underestimation of risk in the majority of patients, whereas the alignment between predicted and observed risks was better for the majority of individuals for the XGBoost and neural network models.

Model	Harrell's C statistic	Calibration slope	Calibration-in-the-large
Cox model	0.782 (0.761 to 0.803) [0.712 to 0.853]	0.962 (0.909 to 1.014) [0.797 to 1.126]	-0.038 (-0.091 to 0.014) [-0.203 to 0.126]
Competing risks regression*	0.711 (0.693 to 0.728) [0.662 to 0.760]	0.573 (0.555 to 0.592) [0.512 to 0.635]	-0.427 (-0.445 to -0.408) [-0.488 to -0.365]
XGBoost	0.748 (0.733 to 0.763) [0.699 to 0.797]	1.068 (1.049 to 1.088) [1.010 to 1.126]	0.068 (0.049 to 0.088) [0.010 to 0.126]
Neural network	0.762 (0.745 to 0.779) [0.707 to 0.817]	1.038 (1.005 to 1.070) [0.933 to 1.142]	0.038 (0.005 to 0.070) [-0.067 to 0.142]
Model	**Royston & Sauerbrei's D Statistic**	**Royston & Sauerbrei's R^2**	
Cox model	1.502 (1.368 to 1.636) [1.060 to 1.943]	0.349 (0.309 to 0.389) [0.217 to 0.481]	

Table 3.9. Summary performance metrics for the 4 models developed to predict 10-year risks of incident breast cancer diagnosis. Royston & Sauerbrei's D Statistic and R^2 are not estimable for the competing risks regression or machine learning models. * = Harrell's C estimated using inverse probability of censoring weighting for the competing risks regression model.

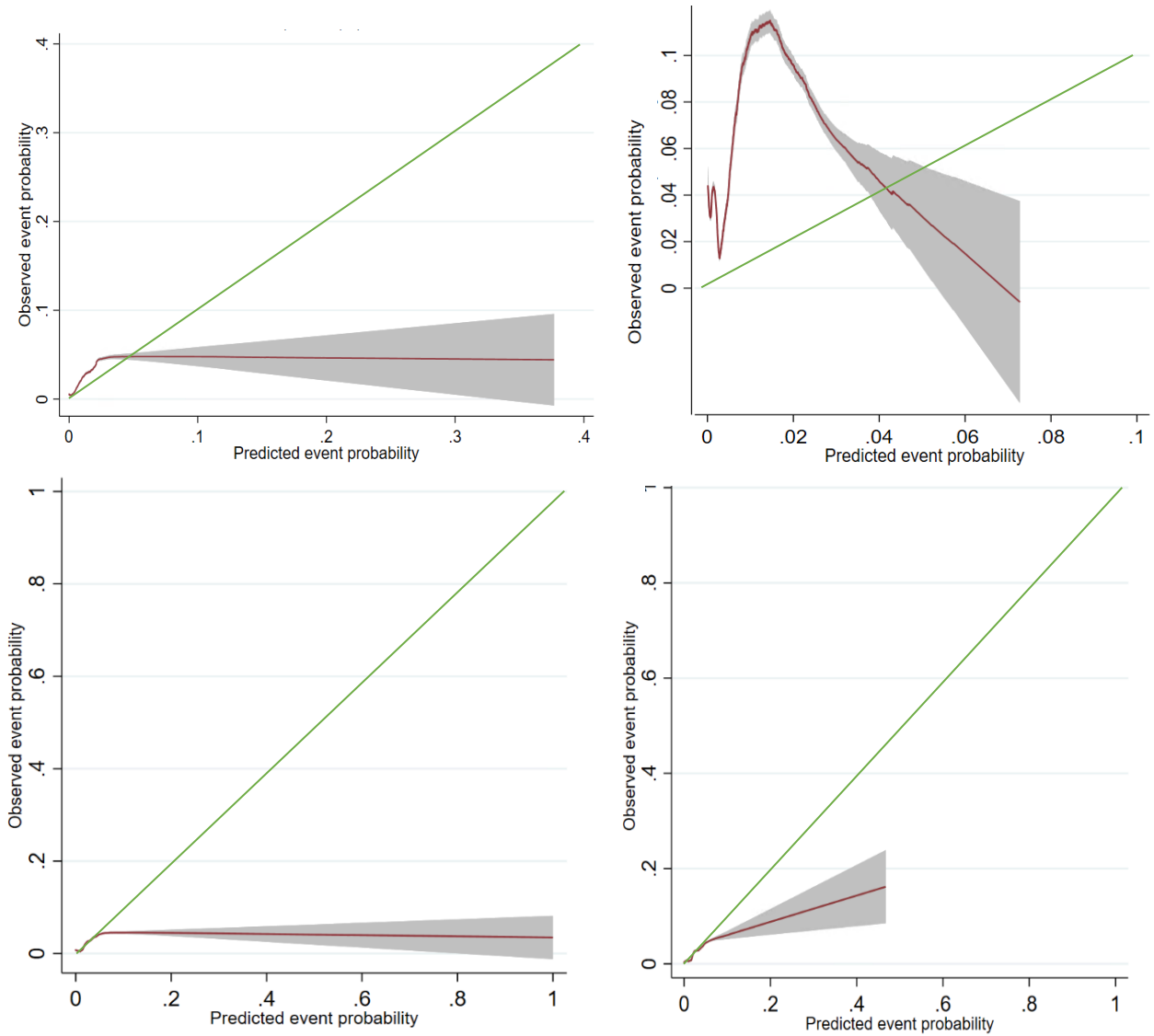

Figure 3.4. Smoothed calibration plots for the final 4 models developed to predict 10-year incident breast cancer risks: top left – Cox model; top right – competing risks regression; bottom left – XGBoost; bottom right – neural network. Scaling for axes is not unform due to the wide ranges in predicted risks generated by each model in the period 2 data.

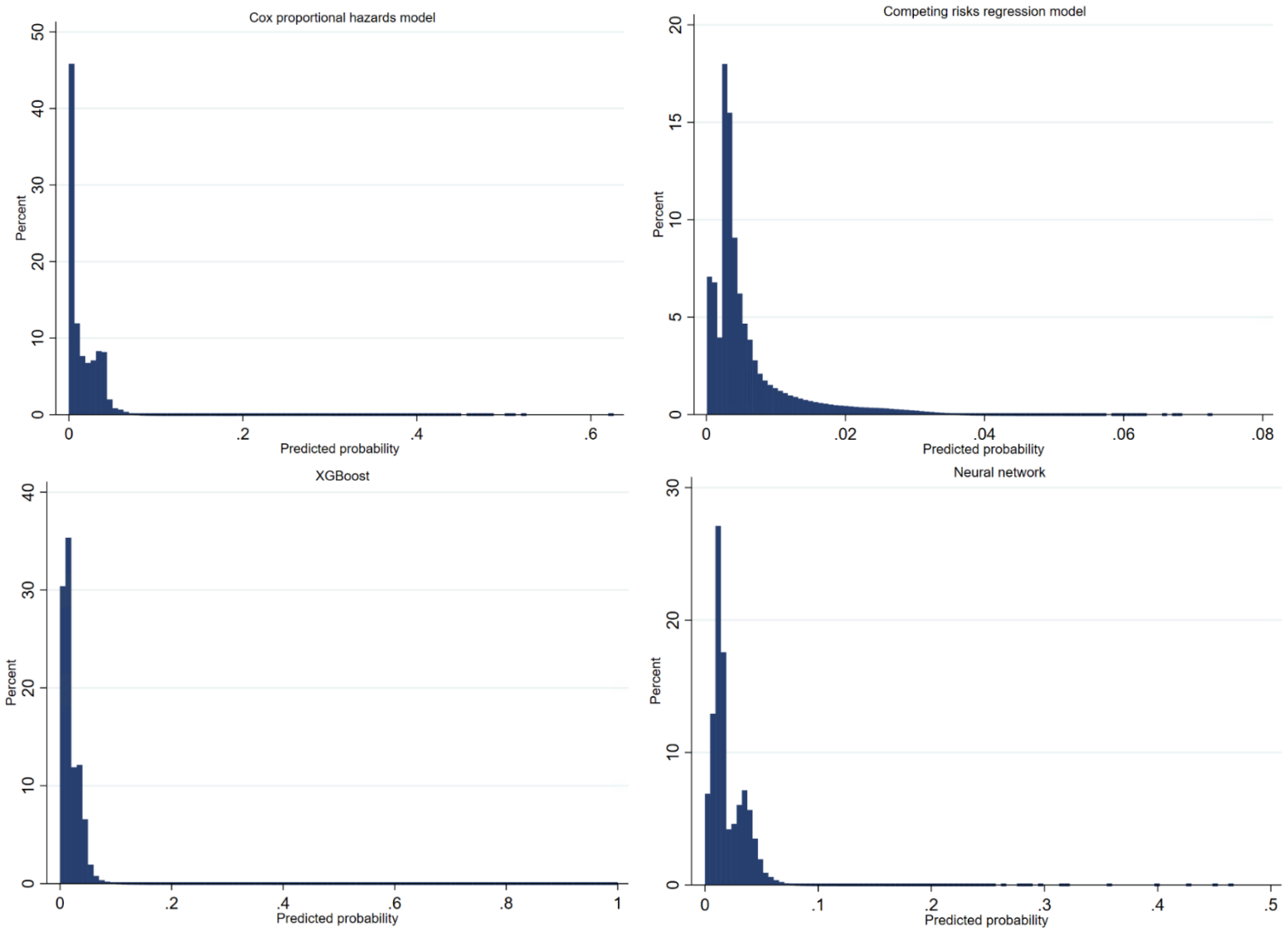

Figure 3.5. Histograms displaying the distribution of predicted risks (probabilities) from all 4 models. The predictions are those generated during internal-external cross-validation. 60,589 patients (1.11%) had a risk of >6% estimated by the Cox model, 95,941 (1.75%) had a risk of >6% estimated by the XGBoost model, and 55,892 (1.02%) had a risk of >6% estimated by the neural network.

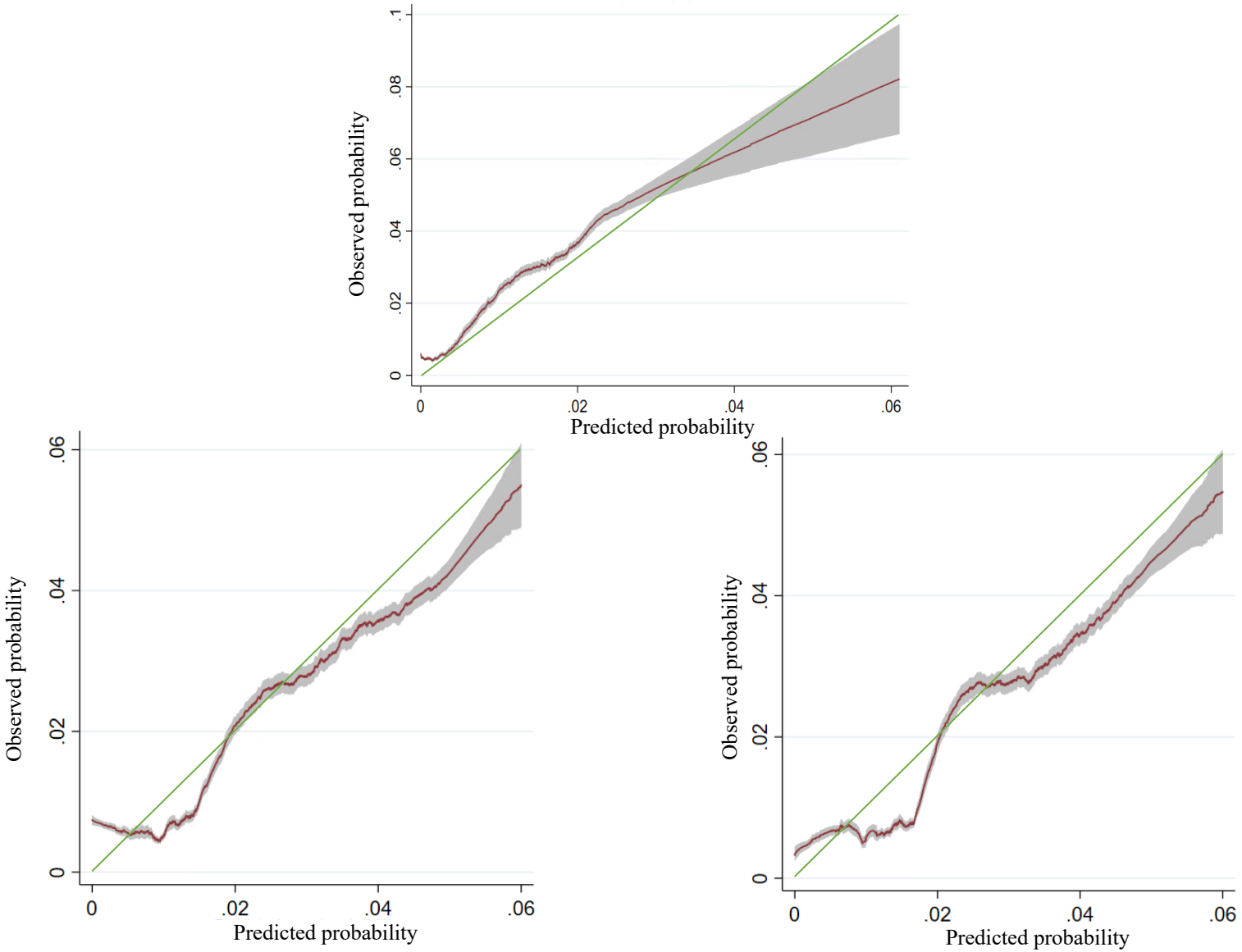

Figure 3.6. Smoothed calibration curves for 3 models. Top = Cox proportional hazards model, bottom left = XGBoost, bottom right = neural network. The green diagonal line represents the ideal. Curves are plotted up until predicted risks of 6% for these risk models for ease of visualisation.

Model evaluation – performance heterogeneity

Figures 3.7 and **3.8** summarise the regional and overall meta-estimates (with confidence and prediction intervals) estimates of Harrell's C and calibration slope for all final models, assessed during the IECV process. Royston & Sauerbrei's D and R^2 statistics are only estimable for the Cox proportional hazards model and are displayed in **Figure 3.9 –** the final model was estimated to account for 34.9% of the variation in time to breast cancer diagnosis (95% CI: 30.9 to 38.9).

Ethnic group performance

The Cox, XGBoost and neural network models had good discrimination in all ethnic groups (point estimates for Harrell's C at 10-years was >0.7, **Tables 3.10** & **3.11**). Ethnic group-specific calibration was generally good for the Cox model. However, there was evidence of slight miscalibration in most non-White ethnic groups for the XGBoost and neural networks, most marked for Bangladeshi women (calibration slope 1.341 [95% CI: 1.130 to 1.552] and 1.354 [95% CI: 1.135 to 1.555]), respectively.

Clinical utility

The highest 5% of predicted risks from all 4 models captured at least 12.75% of all incident breast cancer diagnoses, the highest 25% captured at least 61.37%, and the top 50% captured at least 81.80% (**Table 3.12**). Decision curve analysis demonstrated that the Cox, XGBoost and neural network models were associated with higher net benefit

than ‘treat all’ strategies across a range of thresholds (**Figure 3.10**). The miscalibration of the competing risks model is likely the largest driver of poorer net benefit for this model. In conjunction, these assessments suggest that the Cox, XGBoost and neural network models could have clinical utility in stratifying the female general population for decisions on screening or prevention strategies.

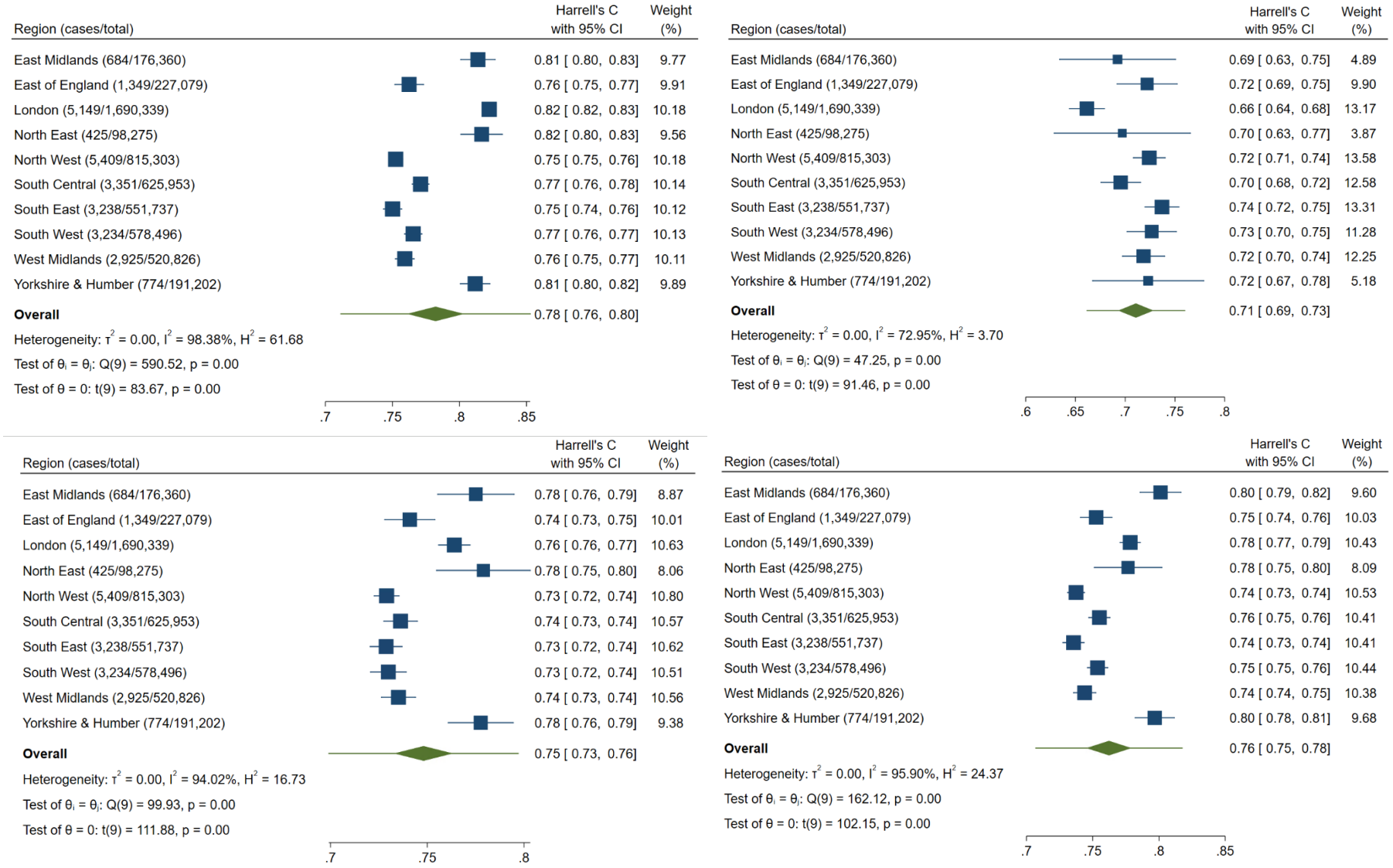

Figure 3.7. Region-level estimates for Harrell's C at 10-years for each of the models – these are estimated using the predictions generated from internal-external cross-validation. Top left = Cox model; top right = competing risks regression; bottom left = XGBoost; bottom right = neural network.

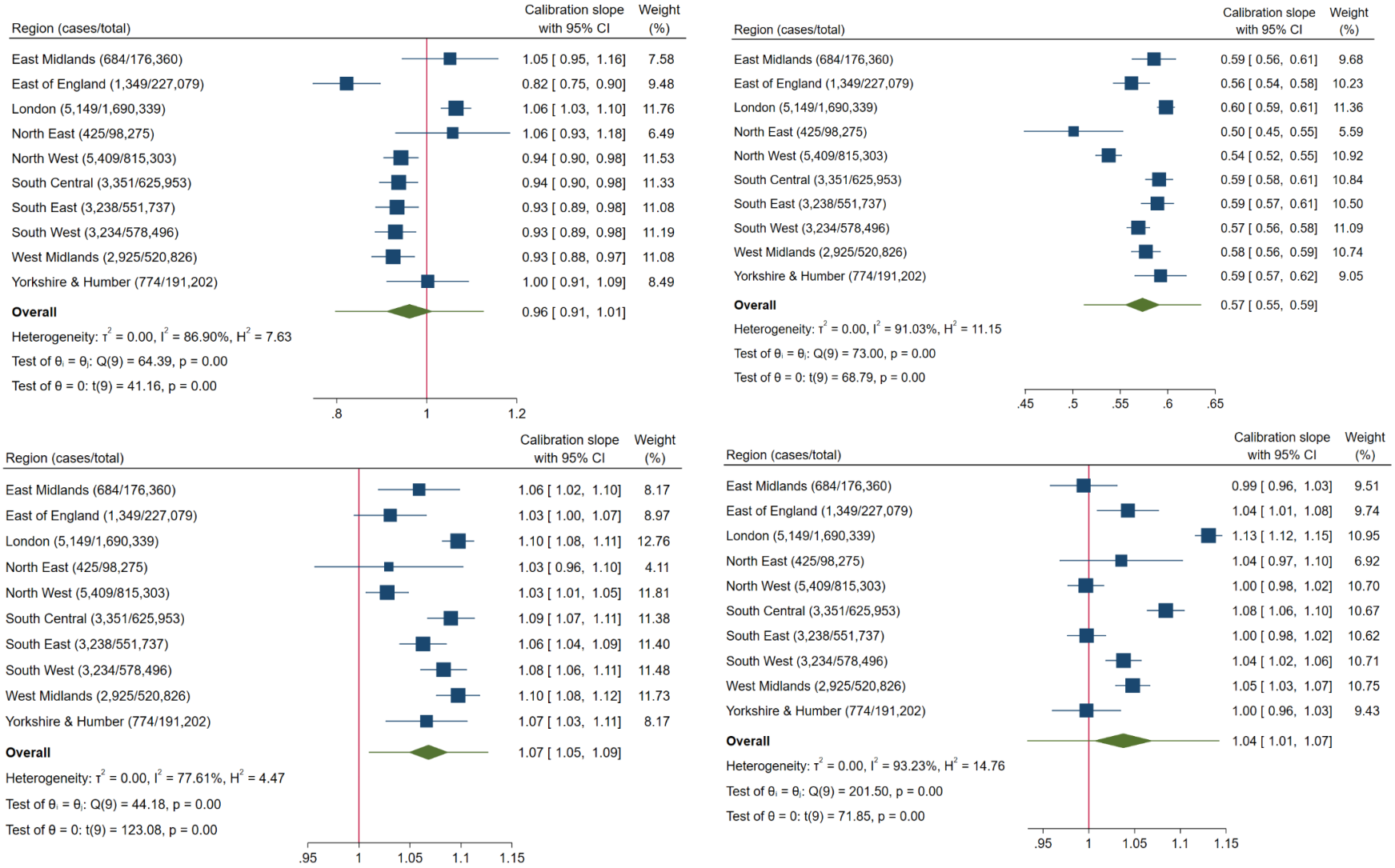

Figure 3.8. Region-level estimates for the calibration slope for each model – these are estimated using the predictions generated from internal-external cross-validation. Top left = Cox model; top right = competing risks regression; bottom left = XGBoost; bottom right = neural network. Red vertical line represents the ideal (value = 1); it is not displayed for the competing risks regression model due to its degree of miscalibration (i.e. far from ideal value).

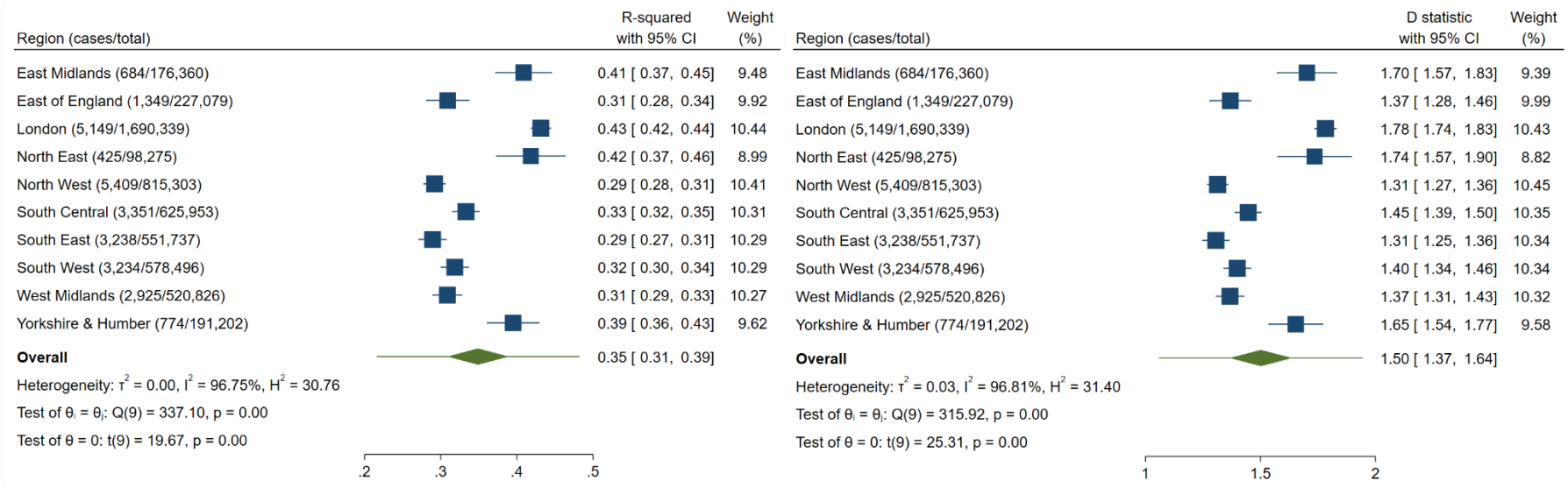

Figure 3.9. Region-level estimates for Royston & Sauerbrei's R^2 (left) and D statistics (right) for the final Cox proportional hazards model – these are estimated using the predictions generated from internal-external cross-validation.

		Cox proportional hazards model			Competing risks regression model		
Ethnic group	Events / denominator	Harrell's C	Calibration slope	Calibration-in-the-large	Harrell's C*	Calibration slope	Calibration-in-the-large
White	18,389 / 3,417,048	0.774 (0.772 to 0.777)	0.945 (0.929 to 0.962)	-0.055 (-0.071 to -0.038)	0.722 (0.714 to 0.730)	0.560 (0.554 to 0.565)	-0.440 (-0.446 to -0.435)
Indian	441 / 160,941	0.822 (0.805 to 0.840)	1.177 (1.046 to 1.309)	0.177 (0.046 to 0.309)	0.711 (0.661 to 0.761)	0.634 (0.600 to 0.667)	-0.366 (-0.400 to -0.333)
Pakistani	254 / 89,626	0.800 (0.777 to 0.823)	1.028 (0.880 to 1.177)	0.028 (-0.120 to 0.177)	0.610 (0.537 to 0.683)	0.707 (0.641 to 0.772)	-0.293 (-0.359 to -0.228)
Bangladeshi	101 / 57,359	0.817 (0.776 to 0.858)	1.021 (0.804 to 1.238)	0.021 (-0.196 to 0.238)	0.556 (0.445 to 0.668)	1.354 (0.325 to 2.382)	0.354 (-0.675 to 1.382)
Other Asian	346 / 123,097	0.802 (0.780 to 0.823)	1.040 (0.898 to 1.183)	0.040 (-0.102 to 0.183)	0.615 (0.559 to 0.671)	0.618 (0.582 to 0.653)	-0.382 (-0.418 to -0.347)
Black Caribbean	337 / 57,270	0.737 (0.715 to 0.759)	0.912 (0.759 to 1.066)	-0.088 (-0.241 to 0.066)	0.687 (0.629 to 0.744)	0.618 (0.565 to 0.672)	-0.382 (-0.435 to -0.328)
Black African	565 / 198,362	0.769 (0.749 to 0.790)	0.852 (0.750 to 0.955)	-0.148 (-0.250 to -0.045)	0.592 (0.536 to 0.648)	0.672 (0.642 to 0.701)	-0.328 (-0.358 to -0.299)
Chinese	140 / 92,940	0.870 (0.842 to 0.898)	1.064 (0.714 to 1.414)	0.064 (-0.286 to 0.414)	0.546 (0.435 to 0.657)	0.605 (0.582 to 0.628)	-0.395 (-0.418 to -0.372)
Other (inc. mixed race, Arab)	576 / 227,249	0.812 (0.798 to 0.827)	1.036 (0.939 to 1.133)	0.036 (-0.061 to 0.133)	0.603 (0.553 to 0.652)	0.614 (0.594 to 0.635)	-0.386 (-0.406 to -0.365)

Table 3.10. Ethnic group-specific performance metrics (with 95% confidence intervals) for each of the regression models developed. * = inverse probability of censoring-weighted Harrell's C. Performance metrics were estimated in the saved predictions resulting from internal-external cross-validation (i.e. in the Period 2 sub-cohort data).

		XGBoost			Neural network		
Ethnic group	Events / denominator	Harrell's C	Calibration slope	Calibration-in-the-large	Harrell's C	Calibration slope	Calibration-in-the-large
White	18,389 / 3,417,048	0.744 (0.741 to 0.747)	1.059 (1.050 to 1.067)	0.059 (0.050 to 0.067)	0.753 (0.750 to 0.757)	1.038 (1.030 to 1.046)	0.038 (0.030 to 0.046)
Indian	441 / 160,941	0.773 (0.748 to 0.799)	1.134 (1.080 to 1.188)	0.134 (0.080 to 0.188)	0.781 (0.756 to 0.806)	1.143 (1.090 to 1.196)	0.143 (0.090 to 0.196)
Pakistani	254 / 89,626	0.751 (0.719 to 0.782)	1.168 (1.084 to 1.251)	0.168 (0.084 to 0.251)	0.743 (0.708 to 0.779)	1.174 (1.095 to 1.253)	0.174 (0.095 to 0.253)
Bangladeshi	101 / 57,359	0.758 (0.694 to 0.821)	1.341 (1.130 to 1.552)	0.341 (0.130 to 0.552)	0.762 (0.704 to 0.819)	1.345 (1.135 to 1.555)	0.345 (0.135 to 0.555)
Other Asian	346 / 123,097	0.748 (0.719 to 0.777)	1.097 (1.032 to 1.163)	0.097 (0.032 to 0.163)	0.744 (0.714 to 0.774)	1.126 (1.067 to 1.184)	0.126 (0.067 to 0.184)
Black Caribbean	337 / 57,270	0.715 (0.688 to 0.741)	1.145 (1.071 to 1.218)	0.145 (0.071 to 0.218)	0.724 (0.698 to 0.750)	1.138 (1.071 to 1.205)	0.138 (0.071 to 0.205)
Black African	565 / 198,362	0.713 (0.686 to 0.740)	1.202 (1.154 to 1.251)	0.202 (0.154 to 0.251)	0.717 (0.691 to 0.744)	1.220 (1.172 to 1.269)	0.220 (0.172 to 0.269)
Chinese	140 / 92,940	0.767 (0.711 to 0.822)	1.065 (1.005 to 1.125)	0.065 (0.005 to 0.125)	0.814 (0.772 to 0.856)	1.087 (1.032 to 1.142)	0.087 (0.032 to 0.142)
Other (inc. mixed race, Arab)	576 / 227,249	0.745 (0.720 to 0.769)	1.096 (1.055 to 1.136)	0.096 (0.055 to 0.136)	0.741 (0.717 to 0.765)	1.108 (1.068 to 1.147)	0.108 (0.068 to 0.147)

Table 3.11. Ethnic group-specific performance metrics (with 95% confidence intervals) for each of the machine learning models developed. * = inverse probability of censoring-weighted Harrell's C. Performance metrics were estimated in the saved predictions resulting from internal-external cross-validation (i.e. in the Period 2 sub-cohort data).

Group of predicted risk (highest)	Cox model		Competing risks regression		XGBoost		Neural network	
	Total incident breast cancer diagnosis in risk group	Cumulative % of total incident breast cancers	Total incident breast cancer diagnosis in risk group	Cumulative % of total incident breast cancers	Total incident breast cancer diagnosis in risk group	Cumulative % of total incident breast cancers	Total incident breast cancer diagnosis in risk group	Cumulative % of total incident breast cancers
1%	974	3.67%	716	2.70%	926	3.49%	1,076	4.05%
2%	1,939	7.31%	1,330	5.01%	1,838	6.93%	2,033	7.66%
3%	2,805	10.57%	1,983	7.47%	2,698	10.17%	2,908	10.96%
4%	3,592	13.54%	2,679	10.09%	3,501	13.19%	3,741	14.10%
5%	4,398	16.57%	3,384	12.75%	4,254	16.03%	4,574	17.24%
10%	8,031	30.26%	7,139	26.90%	7,900	29.77%	8,117	30.58%
15%	11,335	42.71%	10,788	40.65%	11,253	42.40%	11,269	42.46%
20%	14,471	54.53%	13,866	52.25%	14,344	54.05%	14,200	53.51%
25%	17,230	64.93%	16,286	61.37%	17,066	64.31%	16,880	63.61%
50%	25,079	94.50%	21,701	81.80%	23,238	87.57%	23,256	87.63%

Table 3.12. Sensitivity of all 4 models, defined as the percentage of all incident breast cancer cases captured by different cut-offs of the predicted risks distributions. These are estimated using predictions generated during internal-external cross-validation, and therefore the denominator is the number of incident breast cancer diagnoses occurring during the Period 2 sub-cohort.

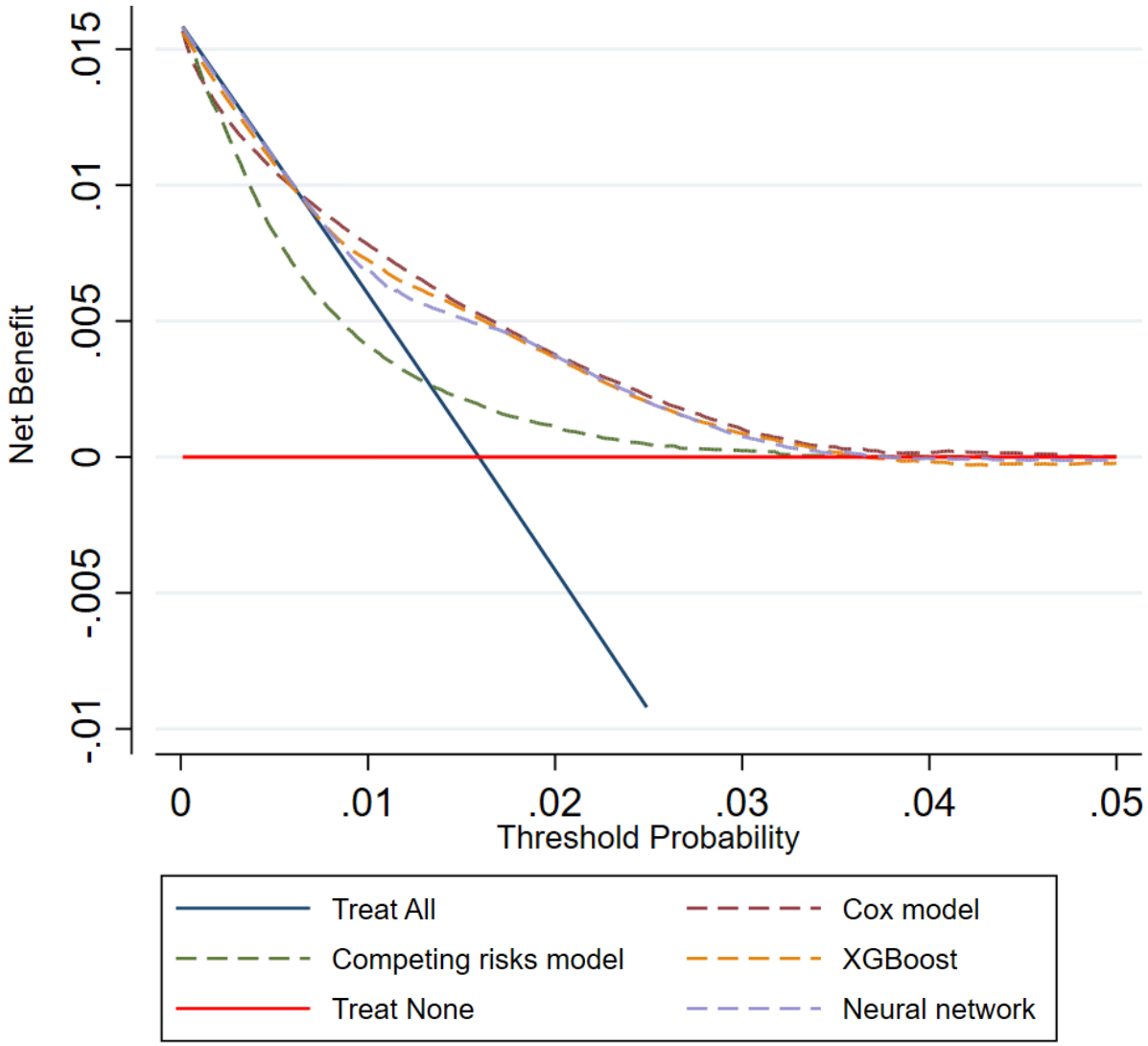

Figure 3.10. Decision curve analysis comparing the net benefit of using all 4 models in terms of their potential effects on clinical decision making. The non-competing risks models generally had similar net benefit associations across the risk spectrum analysed. Whilst the competing risks model had slightly lower discrimination (Harrell's C) than the cause-specific modelling approaches, it is likely that the poor temporal transportability (manifesting as miscalibration) is driving the poor net benefit of this model.

Discussion

This chapter explored four approaches to estimate the 10-year risk of incident breast cancer, including assessment of summary performance metrics, geographical and temporal transportability, and heterogeneity across ethnic groups. Whilst the Cox model had the highest point estimate for Harrell's C, it was miscalibrated for the majority of individuals despite an acceptable summary slope, likely due to combination of under-prediction in lowest risk individuals, and over-prediction in highest risk individuals. The machine learning models had slightly better alignment for most patients as per smoothed calibration plots, and similar net benefit to the Cox model on decision curve analysis.

Whilst there was no single, obvious 'best performing model' in this setting, the IECV strategy identified that the regression model built within a competing risks framework is highly unlikely to be transportable. Poor predictive performance of models on validation may be due to several factors:

1) *Overfitting*

   This is a form of modelling error where the model appears to capture relationships between variables and outcomes well in the derivation data, but performs poorly on other datasets. This is in contrast to 'underfitting', which describes the model inadequately capturing these relationships or the model not including important predictors. Overfitting can be due to the use of datasets of insufficient size to build a reliable model[16], variations in the derivation data that are not present in the validation data, too closely fitting statistical 'noise', or the use of an over-complex algorithm to capture the parameter-outcome associations. Given that the minimum sample size calculations were exceeded with the datasets used to fit these models,

this is unlikely to be a significant driver. Second, the data processing was the same for the overall cohort and period-specific sub-cohorts used for IECV, reducing scope for outliers or similar phenomena to preferentially distort one sub-cohort. Third, the competing risks regression model was formulated as a pseudo-observation based generalised linear model, which is not a complex algorithm structure (indeed far less complex than the XGBoost and neural networks which had thousands of parameters). Therefore, overfitting is unlikely to explain the poor transportability of the competing risks model.

2) *Variation in predictor and/or outcome definitions*

Clinical code groups or other methods used to define measurements and variables for a model may be different between development and validation. For example, use of different Read or SNOMED codes to define breast cancer diagnoses between two datasets may lead to inappropriate ascertainment of cases, with divergent 'ground truth' between datasets. This could be exacerbated in cases where linked datasets are used, such as classifying a baseline exposure based on primary care data, or combinations of primary care and secondary care data. Another example is temporal change in definitions, such as reducing thresholds for diagnosing hypertension over time. However, in the present setting, the same code lists were used to define predictors and outcomes for the data used, meaning this is unlikely to account for the poorer performance of the competing risks model on IECV.

3) *Changes in data availability and linkages*

Related to the second point, in a cohort setting, there can be instances where the data available to define predictors and outcomes vary over time. For example, this could be the occurrence of linkage between primary care and national registry data after follow-up start dates, meaning either predictors or outcomes (or both) are defined differently in time-varying manner. This was not an issue for the linked electronic datasets in this work.

4) *Changes in incidence rates of the outcome of interest*

A major motivation for the use of an IECV framework in this thesis was that trends in incidence and mortality occur over time, in line with changes to treatments, diagnostic pathways and other factors. Changes in risk factor prevalence in the population (e.g. reduced smoking), improved diagnostics (e.g. more sensitive imaging technology), and improved therapy can manipulate incidence or mortality rates, and there may also be changes over time in magnitude of predictor-outcome associations. This is the case for modelling a single risk trajectory, and also when modelling in a competing risks framework. It is arguably more complex in the latter as trends for both the event of interest and any other(s) may be dynamic. A predictive tool based on Cox proportional hazards model comprises a linear predictor and a baseline survival function, which describes the base 'hazard' in the development data. Clearly, temporal trends in incidence may render a model to systematically over- or under-predict in other datasets in which the baseline is different. Another potential issue may be non-proportionality of hazards over time, which may cause a model to perform poorly at different prediction horizons. In this chapter, it was apparent that there were differences in both the outcome of

interest and competing event (other-cause mortality) between the temporally distinct Period 1 and 2 sub-cohorts – this is arguably the likeliest factor driving poorer performance of the competing risks regression model and could also be underpinning miscalibration of the Cox model. Rectifying the performance of a model after external evaluation could involve refitting the model altogether, 'landmarking', or temporal recalibration, which updates the baseline term[17].

Strengths of the approach described in this chapter include the use of linked data sources to aid ascertainment of predictor values and outcomes. The value of the prospective recording of these data points in linked national databases is further enhanced by the lack of selection, recall and respondent biases inherent to the use of routinely collected electronic healthcare data from the representative population. Within the QResearch database anonymised clinical data sharing is on an 'opt-out' basis, rather than self-selecting into a registry such as with UK Biobank[18-20]. Further strengths include the breadth of candidate predictors explored, and the statistical approaches undertaken, including the exploration of non-linearities, interactions, and use of multiple imputation for model development and evaluation. The IECV framework enabled a robust assessment of model performance and emulated a prospective deployment of the final models; in doing so it identified the poor transportability of a competing risks approach for this risk trajectory, and the likely need for ongoing temporal recalibration of cause-specific models predicting breast cancer incidence, should they be deployed at scale[21,22]. This approach was more informative than random splitting, or a single non-random split by region or time, due to the larger sample size available for performance evaluation, notably for numerically smaller population sub-groups. The adaptation of XGBoost and neural network approaches using pseudo-observations modelled risk probabilities

directly, were computationally efficient, and in this scenario were better calibrated than the Cox model (as per smoothed plots) although they may be limited by variation in performance across ethnic groups. Lastly, the use of smoothed calibration plots enabled interrogation of the location and extent of miscalibration across the entire risk spectrum, rather than relying on summary metrics, or approaches where predictions are arbitrarily binned into groups, which can incompletely display or even obscure more complex patterns (**Figure 3.11**).

Limitations include the reliance on clinical practitioners to accurately code predictor variables and outcomes, as free text in clinical notes is not currently evaluable. This, in conjunction with the fact that not all prescriptions will be adhered to, means that there may be misclassification bias for some variables. There may also be bias in recording for variables such as family history. To be recorded, a patient would need to have a relevant family history, this to be noted or ascertained by the practitioner during a consultation, and then a clinical code placed in their electronic healthcare record. Therefore, there may be some misclassification bias where women with more profound family histories are more likely to have this recorded. Conversely, 'positive family history' may be open to interpretation by patients, and some women with less florid family histories (e.g. one second-degree relative affected) may feel that this is significant and volunteer this information during a consultation.

Further, due to the nature of the data, there was no independent adjudication of the outcome(s) of interest. In contrast with other models, this chapter did not include reproductive factor information, genetic data nor mammography-derived data-fields, which may improve model performance. This work did attempt to include information on parity, age at menarche and other facets of reproductive history, but the coding of these in the primary care data were not of sufficient quality with low rates of recording (>90%

missingness) and counterintuitive distributions when data were present. The inability to include genetic data, such as polygenic risk scores, was due to this data not being available in primary or secondary care – this is a subject of interest in a later chapter using a different data source (see **Chapters 6 & 7**).

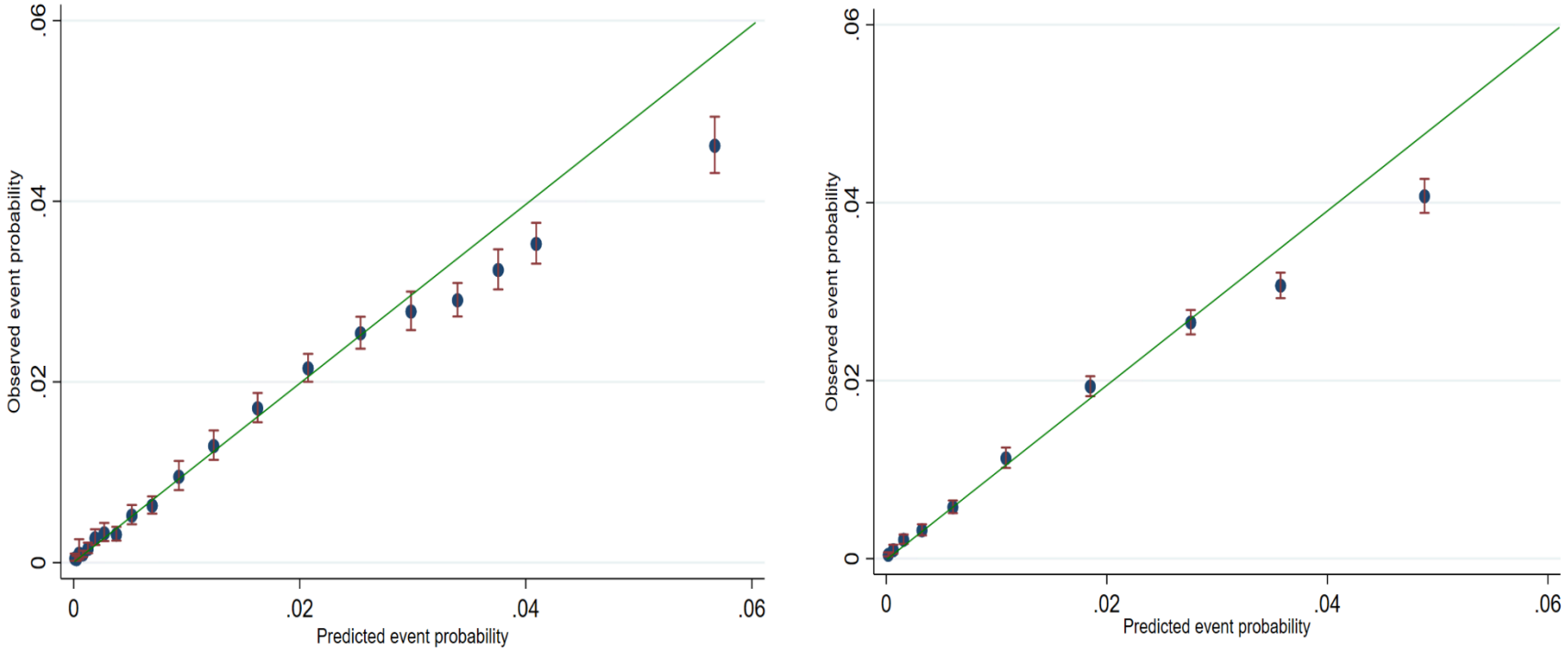

Figure 3.11. Calibration plots for the final Cox proportional hazards generated by binning predicted risks into twentieths (left) and tenths (right). The mean prediction in each sub-group is compared to the 'observed' risk, estimated using the Kaplan-Meier failure function and 95% confidence interval. Whilst the trend towards over-prediction in the very highest risk is visible in the left plot, it appears less profound than on the smoothed plot. The underestimation of risk in lower risk women is not easily visualised. When grouping by tenth of predicted risk, the miscalibration is arguably 'hidden' to an even greater extent.

Lastly, integrated mammography-derived features such as breast density or higher-level abstractions may have predictive relevance[23-27] – however, how much of this is driven by better detection of lesions rather than true longer-term prognostication in women currently without cancer is uncertain, and whilst improved discrimination is reported with some deep learning approaches[24,26,27], their calibration and clinical utility is rarely if ever reported[28]. Most importantly, this work sought to develop models deployable within routine healthcare data infrastructures which may be able to identify high-risk women outside current screening eligibility, whom by definition would be unlikely to have recent mammography imaging available.

This chapter modelled the same trajectory as the QCancer-10 year model[9], but used a larger cohort, considered a broader set of candidate predictors, explored a range of modelling approaches, and used a different validation framework to assess model performance. The final Cox and machine learning models in this chapter do comprise some different predictors – whilst a comparison of these models with QCancer was not possible in this data extract due to overlapping patient records, this is an aspect explored later in this thesis (**Chapter 6 & 7**).

The risk trajectory modelled in this chapter is the standard approach considered for risk-based screening – whilst it may not necessarily be the optimal one to inform strategies to reduce mortality (see **Chapter 1**), it was included in this thesis for provision of a standard approach as a benchmark. This thesis now turns to consider the results from the development and validation of models that predict 10-year risk of mortality after a breast cancer is diagnosed.

Chapter references

1. Clift AK, Dodwell D, Lord S, et al. The current status of risk-stratified breast screening. *Br J Cancer* 2022; **126**(4): 533-50.

2. 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**(11): 687-705.

3. Barlow WE, White E, Ballard-Barbash R, et al. Prospective breast cancer risk prediction model for women undergoing screening mammography. *J Natl Cancer Inst* 2006; **98**(17): 1204-14.

4. Tice JA, Bissell MCS, Miglioretti DL, et al. Validation of the breast cancer surveillance consortium model of breast cancer risk. *Breast Cancer Res Treat* 2019; **175**(2): 519-23.

5. Esserman LJ, Study W, Athena I. The WISDOM Study: breaking the deadlock in the breast cancer screening debate. *NPJ Breast Cancer* 2017; **3**: 34.

6. Louro J, Posso M, Hilton Boon M, et al. A systematic review and quality assessment of individualised breast cancer risk prediction models. *Br J Cancer* 2019; **121**(1): 76-85.

7. Zheng Y, Li J, Wu Z, et al. Risk prediction models for breast cancer: a systematic review. *BMJ Open* 2022; **12**(7): e055398.

8. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Ann Intern Med* 2019; **170**(1): W1-W33.

9. 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**(3): e007825.

10. 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**(5): 489-97.

11. Terry MB, Liao Y, Whittemore AS, et al. 10-year performance of four models of breast cancer risk: a validation study. *Lancet Oncol* 2019; **20**(4): 504-17.

12. Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). *Ann Intern Med* 2015; **162**(10): 735-6.

13. Dhiman P, Ma J, Andaur Navarro CL, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. *BMC Med Res Methodol* 2022; **22**(1): 101.

14. Andersen PK, Perme MP. Pseudo-observations in survival analysis. *Stat Methods Med Res* 2010; **19**(1): 71-99.

15. Klein JP, Andersen PK. Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics* 2005; **61**(1): 223-9.

16. Riley RD, Collins GS, Ensor J, et al. Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome. *Stat Med* 2022; **41**(7): 1280-95.

17. Booth S, Riley RD, Ensor J, et al. Temporal recalibration for improving prognostic model development and risk predictions in settings where survival is improving over time. *Int J Epidemiol* 2020; **49**(4): 1316-25.

18. Collins R. What makes UK Biobank special? *Lancet* 2012; **379**(9822): 1173-4.

19. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* 2017; **186**(9): 1026-34.

20. Tyrrell J, Zheng J, Beaumont R, et al. Genetic predictors of participation in optional components of UK Biobank. *Nat Commun* 2021; **12**(1): 886.

21. Akbarov A, Williams R, Brown B, et al. A Two-stage Dynamic Model to Enable Updating of Clinical Risk Prediction from Longitudinal Health Record Data: Illustrated with Kidney Function. *Stud Health Technol Inform* 2015; **216**: 696-700.

22. Jenkins DA, Martin GP, Sperrin M, et al. Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagn Progn Res* 2021; **5**(1): 1.

23. Brentnall AR, Cuzick J, Buist DSM, et al. Long-term Accuracy of Breast Cancer Risk Assessment Combining Classic Risk Factors and Breast Density. *JAMA Oncol* 2018; **4**(9): e180174.

24. Yala A, Lehman C, Schuster T, et al. A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. *Radiology* 2019; **292**(1): 60-6.

25. Yala A, Mikhael PG, Lehman C, et al. Optimizing risk-based breast cancer screening policies with reinforcement learning. *Nat Med* 2022; **28**(1): 136-43.

26. Yala A, Mikhael PG, Strand F, et al. Multi-Institutional Validation of a Mammography-Based Breast Cancer Risk Model. *J Clin Oncol* 2022; **40**(16): 1732-40.

27. Yala A, Mikhael PG, Strand F, et al. Toward robust mammography-based models for breast cancer risk. *Sci Transl Med* 2021; **13**(578).

28. Lehman CD, Mercaldo S, Lamb LR, et al. Deep Learning vs Traditional Breast Cancer Risk Models to Support Risk-Based Mammography Screening. *J Natl Cancer Inst* 2022; **114**(10): 1355-63.

Chapter Four

Prediction model development and evaluation – 10-year risk of breast cancer mortality following breast cancer diagnosis

Parts of this chapter have been peer reviewed and published previously elsewhere:

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

Summary

Following the development and evaluation of clinical prediction models that estimate incident breast cancer diagnosis risk in **Chapter 3**, this thesis turns to considering approaches to prognostication of women with breast cancer at the point of diagnosis. This chapter summarises the applicability and limitations of current clinical prediction models for breast cancer prognostication, outlines the rationale for developing a novel set of models for this outcome, and then details the results from four analytical approaches using QResearch and linked databases.

Introduction

Accurately estimating the prognosis of women with breast cancer at the point of diagnosis has utility in selecting therapy and could be useful in stratifying follow-up strategies or identifying individuals suitable for clinical trials. Further, evidence from recent simulation studies suggests that adjusting analyses for accurately estimated prognostic scores could improve trial power or make smaller trials possible[1].

Tools such as PREDICT Breast[2,3] and the Nottingham Prognostic Index[4,5] have been or are actively used by clinicians caring for women with early-stage, surgically-treated breast cancer for the purposes of prognostication and to inform selection of adjuvant (post-surgical) treatment. However, these are inherently limited to sub-groups of women with breast cancer, and models that perform well in breast cancers of any type could have broader ranging clinical utility. A systematic review by Phung, et al. appraised original studies published until January 2017 using a modification of the Quality in Prognosis Studies (QUIPS) tool[6], identifying 58 clinical prediction models for early breast cancer from 96 articles – 28 of these predicted mortality[7]. Generally, model performance on external validation was variable with many deemed sub-optimal and differences across different sub-groups were noted, such as reduced performance in patients at the ranges of age or in groups known to be at 'high risk', and the studies that assessed clinical effectiveness typically used simple metrics pertaining to stratifying patients into two 'high' or 'low' risk groups[7]. Recent versions of the PREDICT model were deemed to be at low risk of bias, but sub-optimal model performance has been noted in external validation studies[8]. For example, evaluation in a Dutch cohort of 2,710 patients under the age of 50 years showed model calibration was poor in the lowest and highest risk patients, and another study found that PREDICT under-predicted risk in older (75+ years)

patients[9], those with T3 tumours and those treated with endocrine therapy or chemotherapy[10]. These results demonstrate the need for robust evaluation to not only establish overall performance, but also to search for scenarios where models may not perform well to identify the risks of inappropriate clinical decision making.

A more recent systematic review by Hueting, et al. covering literature published between January 2010 and December 2020 identified 922 breast cancer clinical prediction models, 417 of which (45%) predicted mortality after diagnosis[8]. Of these, 316 models predicted overall survival with an average C-index of 0.74. Using the PROBAST tool[11] rather than adapting a checklist for prognostic factor studies as done by Phung, et al.[7], the authors identified that 95% of all models were at high risk of bias, most had methodological limitations during development and/or validation, and were not reported transparently[8]. Of 27 models that were judged to be at low risk of bias, only one had the intended use of being applied to estimate breast cancer-specific mortality in women with disease of any stage[12]. However, this study by Paredes-Aracil, et al. had several methodological limitations that were not fully recognised in the aforementioned systematic review and appraisal. First, by applying *post-hoc* the minimum sample size calculation approaches of Riley, et al. [13] for model development (published after the primary study), the minimum sample size can be estimated to be 802, versus the 287 women used for modelling by Paredes-Aracil, et al[12] (**Appendix 2**). Second, the modelling process involved dichotomisation of key predictors such as age, which is advised against in best practice[11]. Third, the study examined over 35,000 models developed with different combinations of predictors – this was not clearly accounted for in model selection, and raises concerns over data dredging or the risk of optimism inherent to selecting the best result from a very large number of tests[12]. Therefore, there are no currently available models that reliably

predict breast cancer mortality in women with breast cancer of any stage[8] – this chapter details the results of exploring four modelling approaches to do so.

Study population

In total, 141,765 women aged 20 to 97 years of age at the time of breast cancer diagnosis were included in the final analysis cohort for this sub-study. These were derived from the original cohort from QResearch described in **Chapter 2** – this originally identified women aged 20-90 years at date of entry; of these, women that had a recorded diagnosis of breast cancer were identified. Follow-up was from date of breast cancer diagnosis (earliest date of any recorded breast cancer code in the primary care, HES, or cancer registry) until the date of death, censoring, or study end. Through follow-up (median 4.16 years, interquartile range 1.76 to 8.26 years), 21,688 women had breast cancer-related deaths, and there were 11,454 other-cause deaths; 25,868 (18.25%) had 10 years of follow-up or more. The characteristics of this cohort overall, and by period are summarised in **Table 4.1**. After restricting follow-up to 10-years (prediction horizon), there were 20,367 breast cancer-related deaths within 688,564.81 person-years, with a crude mortality rate of 295.79 (95% CI: 291.75 to 299.88) per 10,000 person-years. Ethnic group-specific and regional mortality rates are summarised in **Table 4.2** and **Table 4.3**, respectively.

On forming the temporally distinct Period 1 and Period 2 sub-cohorts, and truncating follow-up accordingly for IECV, there were 7,551 breast cancer-related deaths within 211,006.95 person-years (12.71%, crude mortality rate 357.96 per 10,000 person-years [95% CI: 349.87 to 366.02]) with 7,554 other-cause deaths (12.72%) in Period 1; and 8,808 breast cancer-related deaths within 297,066.74 person-years (10.69%, crude

mortality rate 296.50 per 10,000 person-years [95% CI: 290.37 to 302.76) and 3,673 other-cause deaths (4.46%) in Period 2.

Model development

Non-linear fractional polynomial terms were selected for age and BMI for both the Cox proportional hazards model and the competing risks regression model (**Figure 4.1**). Breast cancer treatments within the first year of diagnosis were included as auxiliary variables in the imputation model – to avoid 'information leakage', and as the modelling was not performed under a treatment selection/counterfactual framework, these variables were not included as predictors in the final prediction models. In the final modelling, interactions were considered between age and family history of breast cancer, as well as age and BMI.

The final Cox proportional hazards model is reported in full in **Table 4.4.** The final competing risks regression model is reported in full in **Table 4.5**. **Table 4.6** summarises the hyperparameter tuning spaces searched, and the final selected hyperparameter values for the XGBoost and neural network models, which were built in a competing risks framework.

Parameter	Category	Overall study cohort (Column %)	Period 1 sub-cohort (1st Jan 2000-31st Dec 2009) (Column %)	Period 2 sub-cohort (1st Jan 2010 -31st Dec 2020) (Column %)
Total individuals		141,765	59,385	82,380
Age at diagnosis	Mean (SD) Median (IQR)	63.12 (14.31) 62.62 (52.03-73.64)	63.07 (14.35) 62.35 (52.16-74.19)	63.15 (14.28) 62.90 (51.94-73.30)
BMI at diagnosis	Mean (SD) Median (IQR) Not recorded	27.18 (5.55) 26.2 (23.1-30.3) 22,288 (15.72)	26.68 (5.27) 25.8 (22.9-29.6) 14,066 (23.69)	27.48 (5.69) 26.5 (23.3-30.7) 8,222 (9.98)
Townsend deprivation score	Mean (SD)	-0.58 (2.95)	-0.59 (2.96)	-0.57 (2.94)
Ethnicity	White	92,333 (65.13)	33,473 (56.37)	58,860 (71.45)
	Indian	1,666 (1.18)	509 (0.86)	1,157 (1.40)
	Pakistani	866 (0.61)	232 (0.39)	634 (0.77)
	Bangladeshi	317 (0.22)	66 (0.11)	251 (0.30)
	Other Asian	1,019 (0.72)	258 (0.43)	761 (0.92)
	Caribbean	1,320 (0.93)	411 (0.69)	909 (1.10)
	Black African	1,096 (0.77)	242 (0.41)	854 (1.04)
	Chinese	389 (0.27)	104 (0.18)	285 (0.35)
	Other ethnic group (including Arab, mixed race)	1,770 (1.25)	396 (0.67)	1,374 (1.67)
	Not recorded	40,989 (28.91)	23,694 (39.90)	17,295 (20.99)
Smoking status	Non-smoker	80,542 (56.81)	31,744 (53.45)	48,798 (59.24)
	Ex-smoker	31,271 (22.06)	10,999 (18.52)	20,272 (24.61)
	Light smoker (1-9/day)	14,211 (10.02)	6,389 (10.76)	7,822 (9.50)
	Moderate smoker (10-19/day)	3,451 (2.43)	1,525 (2.57)	1,926 (2.34)
	Heavy smoker (20+/day)	1,826 (1.29)	939 (1.58)	887 (1.08)
	Not recorded	10,464 (7.38)	7,789 (13.12)	2,675 (3.25)
Alcohol intake	Non-drinker	76,495 (53.96)	29,181 (49.14)	47,314 (57.43)
	Trivial (<1u/day)	24,711 (17.43)	10,115 (17.03)	14,596 (17.72)
	Light (1-2u/day)	10,123 (7.14)	3,761 (6.33)	6,362 (7.72)
	Moderate (3-6u/day)	6,585 (4.65)	2,476 (4.17)	4,109 (4.99)
	Heavy (7-9u/day)	276 (0.19)	86 (0.14)	190 (0.23)
	Very heavy (>9u/day)	166 (0.12)	24 (0.04)	142 (0.17)
	Not recorded	23,409 (16.51)	13,742 (23.14)	9,667 (11.73)

Cancer grade	Well differentiated	17,439 (12.30)	8,000 (13.47)	9,439 (11.46)
	Moderately differentiated	54,075 (38.14)	21,868 (36.82)	32,207 (39.10)
	Poorly or undifferentiated	34,405 (24.27)	15,153 (25.52)	19,252 (23.37)
	Not recorded	35,846 (25.29)	14,364 (24.19)	21,482 (26.08)
Cancer stage	Stage 1	33,444 (23.59)	10,021 (16.87)	23,423 (28.43)
	Stage 2	29,731 (20.97)	8,258 (13.91)	21,473 (26.07)
	Stage 3	6,358 (4.48)	1,410 (2.37)	4,948 (6.01)
	Stage 4	4,057 (2.86)	1,046 (1.76)	3,011 (3.66)
	Not recorded	68,175 (48.09)	38,650 (65.08)	29,525 (35.84)
Route to cancer diagnosis	Emergency presentation	2,893 (2.04)	1,095 (1.84)	1,798 (2.18)
	GP routine referral	7,699 (5.43)	3,571 (6.01)	4,128 (5.01)
	Inpatient elective	147 (0.10)	76 (0.13)	71 (0.09)
	Other outpatient	1,702 (1.20)	841 (1.42)	861 (1.05)
	Screening	21,149 (14.92)	6,636 (11.17)	14,513 (17.62)
	Two-week wait	36,438 (25.70)	10,610 (17.87)	25,828 (31.35)
	Not recorded	71,737 (50.60)	36,556 (61.56)	35,181 (42.71)
PR status	Negative	9,297 (6.56)	1,366 (2.30)	7,931 (9.63)
	Positive	20,210 (14.26)	2,302 (3.88)	17,908 (21.74)
	Not recorded	112,258 (79.19)	55,717 (93.82)	56,541 (68.63)
HER2 status	Negative	41,571 (29.32)	3,644 (6.14)	37,927 (46.04)
	Positive	7,239 (5.11)	788 (1.33)	6,451 (7.83)
	Not recorded	92,955 (65.57)	54,953 (92.54)	38,002 (46.13)
ER status	Negative	7,930 (5.59)	1,349 (2.27)	6,581 (7.99)
	Positive	44,696 (31.53)	6,682 (11.25)	38,014 (46.14)
	Not recorded	89,139 (62.88)	51,354 (86.48)	37,785 (45.87)
All 3 of PR/ER/HER2 status recorded	Yes	23,795 (16.78)	2,447 (4.12)	21,348 (25.91)
Mastectomy		40,789 (28.77)	17,939 (30.21)	22,850 (27.74)
Other breast surgery		69,584 (49.08)	25,740 (43.34)	43,844 (53.22)
Chemotherapy		38,709 (27.31)	12,823 (21.59)	25,886 (31.42)
Radiotherapy		30,275 (21.36)	2,696 (4.54)	27,579 (33.48)
Family history of breast cancer		6,315 (4.45)	1,916 (3.23)	4,399 (5.34)
Type 1 diabetes mellitus		179 (0.13)	82 (0.14)	97 (0.12)
Type 2 diabetes mellitus		11,410 (8.05)	3,824 (6.44)	7,586 (9.21)
Vasculitis		2,308 (1.63)	802 (1.35)	1,506 (1.83)
Chronic liver disease or cirrhosis		812 (0.57)	247 (0.42)	565 (0.69)

Chronic kidney disease	None/Stage 1/2	131,961 (93.08)	57,410 (96.67)	74,551 (90.50)
	Stage 3	8,959 (6.32)	1,716 (2.89)	7,243 (8.79)
	Stage 4	498 (0.35)	100 (0.17)	398 (0.48)
	Stage 5 (inc. transplant)	347 (0.24)	159 (0.27)	188 (0.23)
Hypertension		44,259 (31.22)	17,135 (28.85)	27,124 (32.93)
Ischaemic heart disease		8,750 (6.17)	3,764 (6.34)	4,986 (6.05)
Systemic lupus erythematosus		303 (0.21)	101 (0.17)	202 (0.25)
Monoamine oxidase inhibitor use*		40 (0.03)	24 (0.04)	16 (0.02)
Oral contraceptive use*		1,892 (1.33)	923 (1.55)	969 (1.18)
Other antidepressant use*		3,071 (2.17)	647 (1.09)	2,424 (2.94)
Renin-angiotensin axis antagonist use*		27,870 (19.66)	9,743 (16.41)	18,127 (22.00)
Selective serotonin reuptake inhibitor use*		12,629 (8.91)	4,017 (6.76)	8,612 (10.45)
Tricyclic antidepressant use*		9,710 (6.85)	3,908 (6.58)	5,802 (7.04)
Thiazide use*		15,769 (11.12)	7,791 (13.12)	7,978 (9.68)
Angiotensin-converting enzyme inhibitor use*		18,806 (13.27)	7,084 (11.93)	11,722 (14.23)
Anti-psychotic medication use*		2,198 (1.55)	1,021 (1.72)	1,177 (1.43)
Beta-blocker use*		16,809 (11.86)	7,220 (12.16)	9,589 (11.64)
Hormone replacement therapy use (any type)*		8,837 (6.23)	5,346 (9.00)	3,491 (4.24)
Missing data in at least 1 variable (before imputation)		131,450 (92.72)	58,482 (98.48)	72,967 (88.57)

Table 4.1. Summary characteristics of the final study cohort overall and separated by temporally distinct sub-cohorts used in internal-external cross-validation. For the purposes of internal-external cross-validation aiming to provide temporally distinct datasets, the follow-up of people

entering the cohort during Period 1 was truncated at the end of that period if necessary (e.g. if they entered in 2002, but died in Period 2015, they would have been censored at end of Period 1); therefore, the numbers of events within 10 years for the sub-cohorts do not equal the overall summary count. BMI = body mass index, IQR = interquartile range, SD -= standard deviation, u = units of alcohol, PR = progesterone receptor, HER2 = human epidermal growth factor 2, ER = oestrogen receptor, GP = general practitioner, * = medication use was defined as at least one prescription in the 6 months preceding breast cancer diagnosis. The median time interval between most recent body mass index record and date of breast cancer diagnosis was 597 days (interquartile range: 193 to 1,661 days). For this endpoint, hormone replacement therapy was handled as a binary categorical feature – this is due to medication-related predictors being defined differently than for the other models, i.e. use within 6 months of diagnosis.

Entire study cohort **Ethnic group**	**Number of person-years**	**Breast cancer deaths**	**Crude mortality rate per 10,000 person-years (95% confidence interval)**
White	480,848.65	10,268	213.91 (209.82 to 218.09)
Indian	7,967.82	170	213.36 (183.58 to 247.97)
Pakistani	3,959.65	88	222.24 (180.34 to 273.88)
Bangladeshi	1,295.81	31	239.23 (168.24 to 340.17)
Other Asian	4,432.10	67	151.17 (118.98 to 192.07)
Black Caribbean	6,065.78	199	328.07 (285.51 to 376.97)
Black African	4,176.94	141	337.57 (286.20 to 398.15)
Chinese	1,776.44	19	106.96 (66.22 to 167.68)
Other, inc. Arab and mixed race	7,458.16	138	185.03 (156.60 to 218.63)
Period 1 sub-cohort*	**Number of person-years**	**Breast cancer deaths**	**Crude mortality rate per 10,000 person-years (95% confidence interval)**
White	131,147.72	1,675	127.72 (121.75 to 133.98)
Indian	1,883.09	41	217.73 (160.32 to 295.70)
Pakistani	757.89	12	158.33 (89.92 to 278.80)
Bangladeshi	206.82	<10	193.40 (72.59 to 515.31)
Other Asian	913.40	12	131.38 (74.61 to 231.35)
Black Caribbean	1,338.78	37	276.37 (200.24 to 381.44)
Black African	631.64	29	459.13 (319.06 to 660.69)
Chinese	359.09	<10	111.39 (41.81 to 296.79)
Other, inc. Arab and mixed race	1,324.16	20	151.04 (97.44 to 234.)
Period 2 sub-cohort	**Number of person-years**	**Breast cancer deaths**	**Crude mortality rate per 10,000 person-years (95% confidence interval)**
White	21,8300.44	5,980	273.93 (267.08 to 280.97)
Indian	4,235.29	92	217.22 (177.08 to 266.47)
Pakistani	2,237.38	57	254.76 (196.51 to 330.28)
Bangladeshi	816.01	24	294.11 (197.14 to 438.90)
Other Asian	2,526.59	43	170.19 (126.22 to 229.48)

Black Caribbean	3,146.55	126	400.44 (336.28 to 476.83)
Black African	2,682.53	99	369.05 (303.07 to 449.41)
Chinese	969.51	<10	92.83 (48.30 to 178.41)
Other, inc. Arab and mixed race	4,618.99	88	190.52 (154.60 to 234.79)

Table 4.2. Ethnic group-specific crude mortality rates in the cohort overall, and separated by temporally distinct period used in internal-external validation (where ethnicity was recorded). * = due to the increased missingness for ethnicity in Period 1 (5,717 out of 7,551 women that had a breast cancer-related death had ethnicity missing) – the majority of these are likely to be White, therefore the true crude mortality rate is higher than estimable here.

Entire study cohort **Region**	**Number of person-years**	**Breast cancer deaths**	**Crude mortality rate per 10,000 person-years (95% confidence interval)**
East Midlands	33,808.54	1,143	338.08 (319.04 to 358.26)
East of England	44,230.81	1,348	304.76 (288.92 to 321.48)
London	103,022.53	2,874	278.97 (268.95 to 289.36)
North East	25,682.71	892	347.32 (325.25 to 370.87)
North West	121,395.95	3,677	302.89 (293.26 to 312.84)
South Central	94,206.03	2,512	266.65 (256.42 to 277.28)
South East	68,119.61	1,917	281.42 (269.10 to 294.30)
South West	85,956.20	2,338	271.99 (261.19 to 283.25)
West Midlands	74,718.15	2,448	327.63 (314.01 to 340.87)
Yorkshire & Humber	37,424.29	1,218	327.63 (314.91 to 340.87)
Total	*688,564.81*	*20,367*	295.79 (291.75 to 299.88)
Period 1 sub-cohort	**Number of person-years**	**Breast cancer deaths**	**Crude mortality rate per 10,000 person-years (95% confidence interval)**
East Midlands	12,702.06	456	358.99 (327.51 to 393.51)
East of England	14,989.59	513	342.24 (313.87 to 373.17)
London	31,548.06	1,080	342.33 (322.52 to 363.37)
North East	8,082.93	307	379.81 (339.62 to 414.77)
North West	33,929.49	1,301	383.44 (363.16 to 404.85)
South Central	27,730.34	960	346.19 (324.97 to 368.80)
South East	19,817.87	685	345.65 (320.71 to 372.53)
South West	26,160.24	858	327.98 (306.75 to 350.68)
West Midlands	23,627.17	892	377.53 (353.55 to 403.14)
Yorkshire & Humber	12,419.20	499	401.80 (368.05 to 438.64)
Total	*211,006.95*	*7,551*	*357.86 (349.87 to 366.02)*
Period 2 sub-cohort	**Number of person-years**	**Breast cancer deaths**	**Crude mortality rate per 10,000 person-years (95% confidence interval)**

East Midlands	11,178.88	464	415.07 (378.97 to 454.61)
East of England	16,713.76	543	324.88 (298.67 to 353.39)
London	46,278.92	1,270	274.42 (259.74 to 289.94)
North East	9,887.09	387	391.42 (354.30 to 432.43)
North West	57,826.99	1,664	287.75 (274.26 to 301.92)
South Central	41,874.18	1,062	253.62 (238.81 to 269.34)
South East	31,040.41	848	273.19 (255.41 to 292.21)
South West	37,130.95	1,054	283.86 (267.23 to 301.53)
West Midlands	30,691.94	1,032	336.24 (316.34 to 357.40)
Yorkshire & Humber	14,443.62	484	335.10 (306.53 to 366.32)
Total	*29,7066.74*	*8,808*	*296.50 (290.37 to 302.76)*

Table 4.3. Regional crude mortality rates for the cohort overall and separated by temporally distinct period used in internal-external validation.

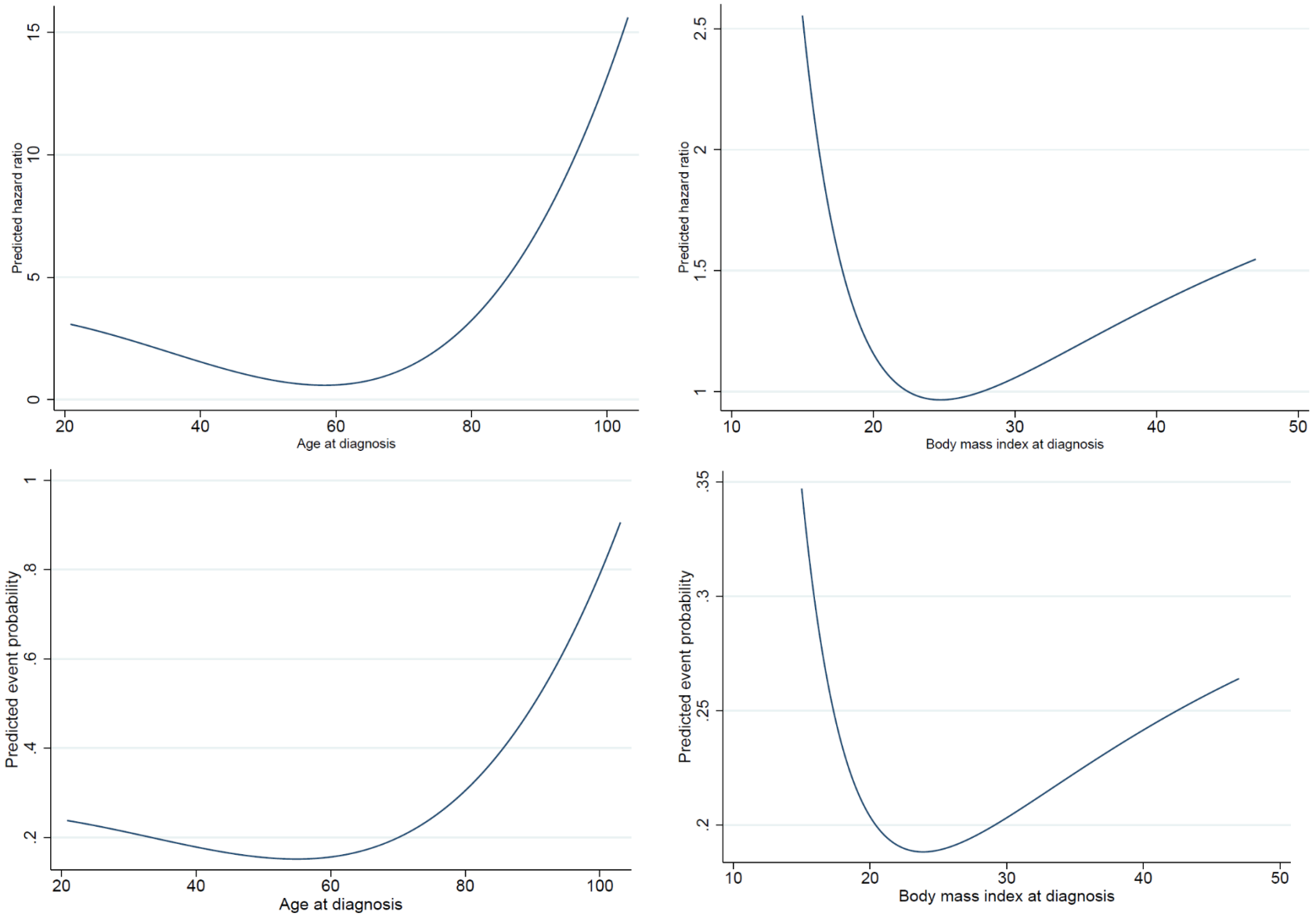

Figure 4.1. Fractional polynomial terms for age and body mass index, for the Cox proportional hazards (top row), and competing risks regression models (bottom row).

Parameter	Description	Coefficient
Age at diagnosis (1st FP term)	X^0.5 - 2.510345299 X = age/10	-2.4200838
Age at diagnosis (2nd FP term)	X^2 - 39.71310575 X = age/10	0.06170782
BMI at diagnosis (1st FP term)	X^-2 - 0.1353749689 X = BMI/10	3.5626427
BMI at diagnosis (2nd FP term)	X^-2*ln(X) - 0.1353551232 X = BMI/10	-8.2391665
Smoking status	Non-smoker (reference)	0
	Ex-smoker	0.07487781
	Light smoker	0.33338149
	Moderate smoker	0.35168628
	Heavy smoker	0.51960507
Route to breast cancer diagnosis	Emergency presentation	1.8266128
	GP referral	0.87391787
	Inpatient elective	1.5714949
	Other outpatient	1.0395208
	Screening (reference)	0
	Two-week wait	0.79912031
Progesterone receptor status	Negative (reference)	0
	Positive	-0.36406436
HER2 status	Negative (reference)	0
	Positive	-0.18971053
Oestrogen receptor status	Negative (reference)	0
	Positive	-0.30338542
Cancer stage at diagnosis	Stage 1 (reference)	0
	Stage 2	1.0091567
	Stage 3	1.8234394
	Stage 4	2.4389817
Cancer grade	Well differentiated (reference)	0
	Moderately differentiated	0.28412054
	Poorly or undifferentiated	0.59900683
Chronic kidney disease	None/Stage 1/2 (reference)	0
	Stage 3	0.06287756
	Stage 4	0.43083109
	Stage 5 (inc. transplant)	0.49398542
Chronic liver disease	None (reference)	0
	Yes	0.316256
Type 2 diabetes mellitus	No (reference)	0
	Yes	0.10158724
ACE inhibitor use	No (reference)	0
(within 6 months prior)	Yes	0.14908664
Renin angiotensin axis inhibitor use	No (reference)	0

(within 6 months prior)	Yes	-0.14476126
Tricyclic antidepressant use	No (reference)	0
(within 6 months prior)	Yes	0.13668533
Selective serotonin reuptake inhibitor use	No (reference)	0
(within 6 months prior)	Yes	0.16313084
Other antidepressant use	No (reference)	0
(within 6 months prior)	Yes	0.18472118
HRT use	No (reference)	0
(within 6 months prior)	Yes	-0.285108
Anti-psychotic use	No (reference)	0
(within 6 months prior)	Yes	0.43001608
Baseline survival function at 10 years		0.9592283

Table 4.4. Final Cox proportional hazards model coefficients and baseline survival term. FP = fractional polynomial; HER2 = human epidermal growth factor receptor 2; ACE = angiotensin-converting enzyme; HRT = hormone replacement therapy.

Parameter	Description	Coefficient
Age at diagnosis (1st FP term)	X – 6.301833523 X = age/10	-0.33729257
Age at diagnosis (2nd FP term)	X^2 - 39.71310575 X = age/10	0.04244611
BMI (1st FP term)	X^-2 - 0.1353749689 X = BMI/10	2.2932369
BMI (2nd FP term)	X^-2*ln(X) - 0.1353551232 X = BMI/10	-5.9806931
Smoking status	Non-smoker (reference)	0
	Ex-smoker	0.04905292
	Light smoker	0.29232912
	Moderate smoker	0.24918172
	Heavy smoker	0.48453944
Route to breast cancer diagnosis	Emergency presentation	1.208285
	GP referral	0.74362281
	Inpatient elective	1.5009607
	Other outpatient	0.9191697
	Screening (reference)	0
	Two-week wait	0.6927579
Progesterone receptor status	Negative (reference)	0
	Positive	-0.40245066
HER2 status	Negative (reference)	0
	Positive	-0.20705725
Oestrogen receptor status	Negative (reference)	0
	Positive	-0.22475207
Cancer stage at diagnosis	Stage 1 (reference)	0
	Stage 2	0.91589078
	Stage 3	1.7065489
	Stage 4	2.3440632
Cancer grade	Well differentiated (reference)	0
	Moderately differentiated	0.29826785
	Poorly or undifferentiated	0.52620672
HRT use	No (reference)	0
(within 6 months prior)	Yes	-0.18103964
Anti-psychotic medication use		0
(within 6 months prior)		0.21301738
Constant		-2.9552694

Table 4.5. Full competing risks regression model and constant term. FP = fractional polynomial; HER2 = human epidermal growth factor receptor 2; HRT = hormone replacement therapy.

Model	Basic architecture	Hyperparameters tuned	Range explored during tuning	Final selected value after tuning
XGBoost	Tree-based booster with GPU_hist	Maximum tree depth	1 to 6	6
		Learning rate (eta)	0.0001 to 0.1	0.073
	Gradient-based sub-sampling	Subsampling proportion	1.1 to 0.5	0.1
		Number of boosting rounds	0 to 500	251
	RMSE evaluation metric	Alpha (regularisation)	0 to 20	18
		Gamma (regularisation)	0 to 20	0
		Lambda (regularisation)	0 to 20	3
		Column sampling by tree	0.1 to 0.8	0.501
		Column sampling by level	0.1 to 0.8	0.518
Neural network	Feed-forward ANN with fully connected layers	Number of hidden layers	1 to 5	2
	26 input nodes	Number of nodes in each hidden layer	26 to 50	30
	ReLU activation functions in hidden layers	Number of epochs	1 to 50	32
	Adam optimiser	Initial learning rate	0.001 to 0.1	0.032
	Single output node with linear activation			
	RMSE loss function			
	Batch size 1024			

Table 4.6. Description of machine learning model architectures and hyperparameter tuning performed. The continuous outcome variables for both models were the jack-knife pseudo-observations for the Aalen-Johansen cumulative incidence function at 10 years. RMSE = root mean squared error, ReLU – rectified linear unit, ANN = artificial neural network. The final neural network model had a total of 1,771 parameters (all trainable).

Model evaluation – overall performance

The summary discrimination and calibration metrics for all 4 models are summarised in **Table 4.7**.

The Cox proportional hazards model had the highest point estimate for meta-analysis pooled Harrell's C at 0.858 (95% CI: 0.853 to 0.864, 95% prediction interval: 0.843 to 0.873), but its confidence and prediction intervals overlapped with those for the competing risks regression model (meta-analysis pooled Harrell's C of 0.849 [95% CI: 0.839 to 0.859, 95% prediction interval 0.821 to 0.876]).

There was a small degree of miscalibration on summary metrics for most models, such as the meta-analysis pooled estimate for the calibration slope of 1.108 (95% CI: 1.079 to 1.138, 95% prediction interval 1.034 to 1.182) for the Cox model.

Calibration plots (**Figure 4.2**) demonstrated good alignment of predicted and observed risks across the predicted risk spectrum for the Cox model. However, there was some overestimation of risk for higher risk women with the competing risks regression above a threshold of approximately 40%, and the two machine learning models demonstrated more complex patterns of miscalibration that were not appreciable with summary measures.

Model evaluation – performance heterogeneity

All 4 final models appeared to have good discrimination across all ethnic groups, although some metrics were estimated with reduced precision due to lower event counts **(Tables 4.8 and 4.9).** The calibration of the models across ethnic groups was generally acceptable but variable, with each model tending towards miscalibration in at least 2 ethnic groups.

The patterns of performance in 10-year age groups were similar – all models appeared to have good discriminatory capability in each age group with the point estimates for Harrell's C for the regression models generally being higher for the corresponding sub-group analyses of the machine learning model performance. Summary calibration metrics were more erratic (**Tables 4.10 & 4.11**)

As the intention was to develop models that reliably predict prognosis in women with breast cancer of any stage, tumour stage-specific performance analyses were undertaken. The general trend herein was that discrimination reduced with increasing stage, and miscalibration occurred in higher stage tumours. Varying extents of miscalibration were observed across tumour stages with the regression models, and all models tended to be less well calibrated for stage IV tumours (**Table 4.12**). This was most pronounced with the XGBoost and neural networks in stage III and stage IV tumours. This likely reflects variation in predictor-outcome associations across stage and baseline rates, and suggests that for optimal model performance, regression models may require stage-specific recalibration, such as using a stage-specific baseline function; for machine learning models, stage-specific models may be required to be developed.

Model	Harrell's C	Calibration slope	Calibration in the large
Cox model	0.858 (0.853 to 0.864) [0.843 to 0.873]	1.108 (1.079 to 1.138) [1.034 to 1.182]	0.108 (0.079 to 0.138) [0.034 to 0.182]
Competing risks model	0.849 (0.839 to 0.859) [0.821 to 0.876]	1.160 (1.064 to 1.255) [0.872 to 1.447]	0.160 (0.064 to 0.255) [-0.218 to 0.447]
XGBoost	0.821 (0.813 to 0.828) [0.805 to 0.837]	1.084 (1.003 to 1.165) [0.842 to 1.326]	0.084 (0.003 to 0.165) [-0.158 to 0.326]
Neural network	0.847 (0.835 to 0.858) [0.816 to 0.878]	1.037 (0.910 to 1.165) [0.624 to 1.451]	0.037 (-0.090 to 0.165) [-0.376 to 0.451]

Table 4.7. Summary performance metrics for all 4 models, estimated using random effects meta-analysis after internal-external cross-validation. Metrics are presented as the point estimate, (confidence interval), and [95% prediction interval]. Harrell's C for the competing risks, XGBoost and neural network models was estimated using inverse probability of censoring weighting.

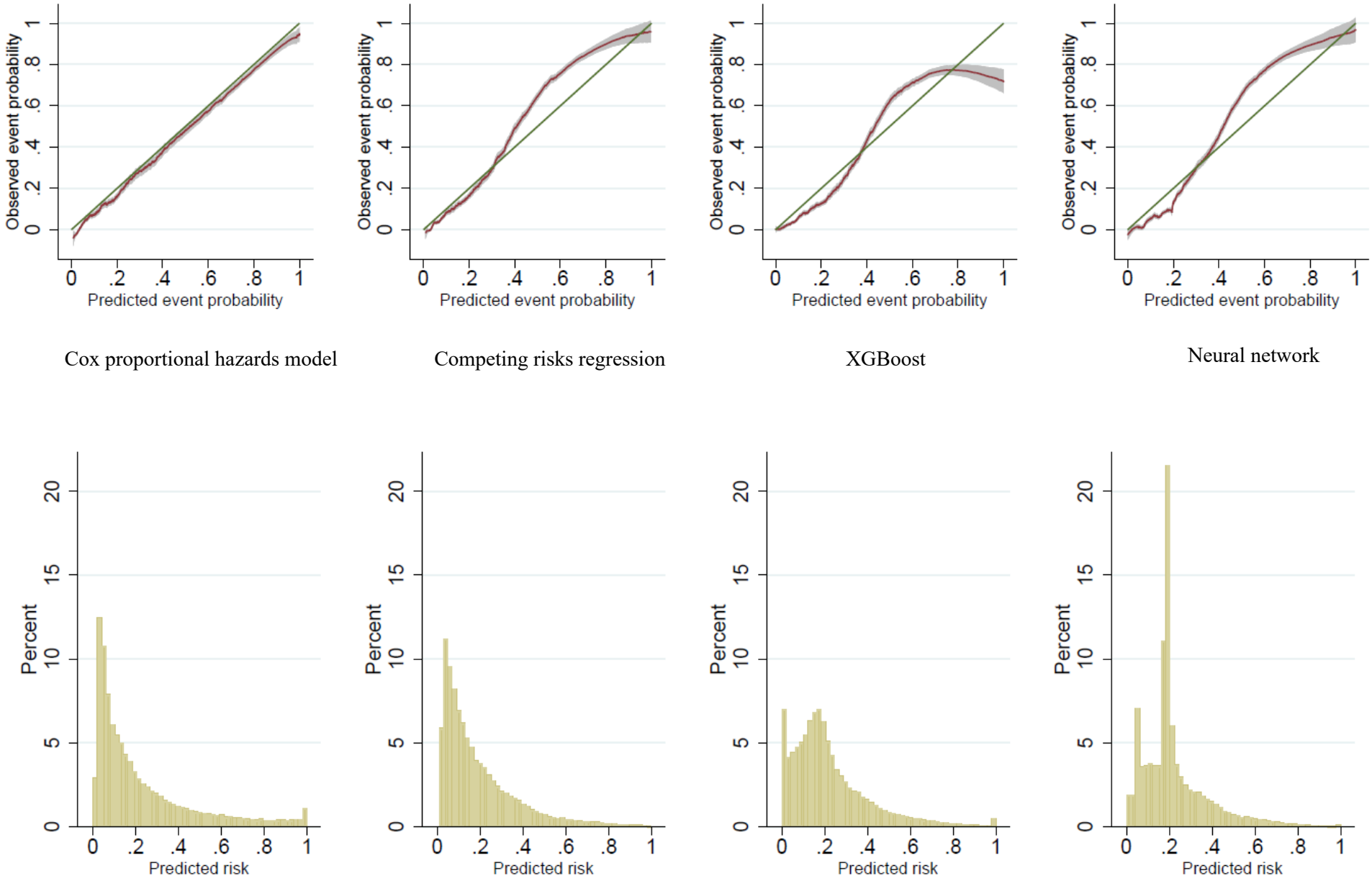

Figure 4.2. Calibration plots and predicted risk distributions for each of the four final models.

Clinical utility

Table 4.13 summarises the sensitivity of each model for breast cancer-related death across several cut-offs of their predicted risks distributions. The overall decision curves demonstrated the superiority of the Cox and competing risks regression models over the XGBoost and neural network models for most threshold probabilities (**Figure 4.3**). When examining clinical utility by stage, the gap in performance of the machine learning models was most pronounced for lower stage tumours (**Figure 4.4**).

Meta-regression

Regional differences in the Harrell's C-index were relatively slight. Meta-analytic pooling estimated an I^2 of 53.14% for the Cox model, with no variables associated with such variation in terms of discrimination. A small degree of inter-region heterogeneity was observed for calibration, with none of this attributable to regional variation in any of the sociodemographic factors examined. It was estimated that 41.33% of the regional variation in the Harrell's C-index for the competing risks regression model was attributable to inter-regional case mix (**Table 4.14**; ethnic diversity and to a lesser extent age were the only sociodemographic factors associated therewith). Regarding calibration slope for the competing risks model, the I^2 was 56.68%, with regional variation in age, deprivation and ethnic diversity associated therewith (total R^2 60.08%). Meta-regression estimated that the leading factor associated with regional variation in discrimination and calibration metrics was regional differences in ethnic diversity (**Table 4.14**).

		Cox proportional hazards model			Competing risks regression model		
Ethnic group	Events / denominator	Harrell's C	Calibration slope	Calibration-in-the-large	Harrell's C*	Calibration slope	Calibration-in-the-large
White	5,980 / 58,860	0.860 (0.855 to 0.864)	1.106 (1.077 to 1.136)	0.106 (0.077 to 0.136)	0.848 (0.840 to 0.856)	1.183 (1.128 to 1.238)	0.184 (0.128 to 0.238)
Indian	92 / 1,157	0.849 (0.810 to 0.888)	1.150 (0.960 to 1.340)	0.150 (-0.040 to 0.340)	0.843 (0.780 to 0.906)	1.367 (1.029 to 1.705)	0.367 (0.029 to 0.705)
Pakistani	57 / 634	0.845 (0.790 to 0.899)	1.277 (0.935 to 1.618)	0.277 (-0.065 to 0.618)	0.812 (0.720 to 0.903)	1.230 (0.821 to 1.639)	0.230 (-0.179 to 0.639)
Bangladeshi	24 / 251	0.794 (0.691 to 0.896)	1.057 (0.571 to 1.543)	0.057 (-0.429 to 0.543)	0.777 (0.645 to 0.909)	1.666 (0.939 to 2.394)	0.666 (-0.061 to 1.394)
Other Asian	43 / 761	0.818 (0.752 to 0.884)	1.105 (0.790 to 1.420)	0.105 (-0.210 to 0.420)	0.849 (0.757 to 0.941)	1.548 (1.104 to 1.993)	0.548 (0.104 to 0.993)
Black Caribbean	126 / 909	0.822 (0.784 to 0.860)	1.097 (0.895 to 1.299)	0.097 (-0.105 to 0.299)	0.853 (0.801 to 0.904)	1.188 (0.901 to 1.475)	0.188 (-0.099 to 0.475)
Black African	99 / 854	0.831 (0.788 to 0.874)	0.996 (0.806 to 1.186)	-0.004 (-0.194 to 0.186)	0.822 (0.760 to 0.883)	1.106 (0.783 to 1.429)	0.106 (-0.217 to 0.429)
Chinese	9 / 285	0.931 (0.839 to 1.000)	1.900 (0.984 to 2.817)	0.900 (-0.016 to 1.817)	0.916 (0.802 to 1.000)	1.940 (1.115 to 2.764)	0.940 (0.115 to 1.764)
Other ethnic group	88 / 1,374	0.821 (0.772 to 0.869)	1.124 (0.910 to 1.339)	0.124 (-0.090 to 0.339)	0.834 (0.769 to 0.898)	1.667 (1.301 to 2.034)	0.667 (0.301 to 1.034)

Table 4.8. Ethnic group-specific regression model performance metrics (with 95% confidence intervals) estimated after internal-external cross-validation in the data from period 2 (2010-2020). Event and denominator counts for each ethnic group are from the 'complete case' data for reference, but performance metrics were calculated using the multiply imputed datasets. * = weighted by the inverse probability of censoring.

		XGBoost			Neural network		
Ethnic group	Events / denominator	Harrell's C*	Calibration slope	Calibration-in-the-large	Harrell's C*	Calibration slope	Calibration-in-the-large
White	5,980 / 58,860	0.819 (0.810 to 0.828)	1.098 (1.058 to 1.139)	0.098 (0.058 to 0.139)	0.797 (0.785 to 0.808)	1.085 (1.053 to 1.118)	0.085 (0.053 to 0.118)
Indian	92 / 1,157	0.839 (0.783 to 0.894)	1.251 (0.943 to 1.560)	0.251 (-0.057 to 0.560)	0.840 (0.770 to 0.910)	1.227 (0.970 to 1.484)	0.227 (-0.030 to 0.484)
Pakistani	57 / 634	0.807 (0.717 to 0.897)	1.227 (0.824 to 1.630)	0.227 (-0.176 to 0.630)	0.732 (0.622 to 0.842)	1.079 (0.781 to 1.377)	0.079 (-0.219 to 0.377)
Bangladeshi	24 / 251	0.722 (0.592 to 0.853)	1.306 (0.791 to 1.820)	0.306 (-0.209 to 0.820)	0.738 (0.597 to 0.879)	1.392 (0.940 to 1.844)	0.392 (-0.060 to 0.844)
Other Asian	43 / 761	0.830 (0.735 to 0.925)	1.507 (1.008 to 2.007)	0.507 (0.008 to 1.007)	0.823 (0.715 to 0.930)	1.429 (1.078 to 1.780)	0.429 (0.078 to 0.780)
Black Caribbean	126 / 909	0.797 (0.727 to 0.868)	1.040 (0.781 to 1.300)	0.040 (-0.219 to 0.300)	0.816 (0.740 to 0.891)	1.053 (0.839 to 1.267)	0.053 (-0.161 to 0.267)
Black African	99 / 854	0.806 (0.745 to 0.866)	1.051 (0.757 to 1.345)	0.051 (-0.243 to 0.345)	0.806 (0.733 to 0.879)	0.989 (0.747 to 1.232)	-0.011 (-0.253 to 0.232)
Chinese	9 / 285	0.912 (0.784 to 1.000)	1.862 (1.017 to 2.708)	0.862 (0.017 to 1.708)	0.931 (0.821 to 1.000)	1.704 (1.067 to 2.341)	0.704 (0.067 to 1.341)
Other ethnic group	88 / 1,374	0.809 (0.753 to 0.865)	1.579 (1.230 to 1.927)	0.579 (0.230 to 0.927)	0.827 (0.754 to 0.900)	1.495 (1.214 to 1.776)	0.495 (0.214 to 0.776)

Table 4.9. Ethnic group-specific machine learning model performance metrics (with 95% confidence intervals) estimated after internal-external cross-validation in the data from period 2 (2010-2020). Event and denominator counts for each ethnic group are from the 'complete case' data for reference, but performance metrics were calculated using the multiply imputed datasets. * = weighted by the inverse probability of censoring.

		Cox proportional hazards model			Competing risks regression model		
Age group	Events / denominator	Harrell's C	Calibration slope	Calibration-in-the-large	Harrell's C*	Calibration slope	Calibration-in-the-large
20-29 years	26 / 317	0.821 (0.719 to 0.906)	1.308 (0.661 to 1.955)	0.308 (-0.339 to 0.955)	0.849 (0.748 to 0.951)	1.553 (0.930 to 2.176)	0.553 (-0.070 to 1.176)
30-39 years	287 / 3,259	0.786 (0.754 to 0.817)	1.196 (1.025 to 1.367)	0.196 (0.025 to 0.367)	0.809 (0.766 to 0.852)	1.317 (1.053 to 1.582)	0.317 (0.053 to 0.582)
40-49 years	865 / 12,398	0.833 (0.818 to 0.848)	1.352 (1.246 to 1.459)	0.352 (0.246 to 0.459)	0.848 (0.828 to 0.867)	1.421 (1.263 to 1.579)	0.421 (0.263 to 0.579)
50-59 years	1,230 / 19,648	0.866 (0.854 to 0.877)	1.264 (1.193 to 1.335)	0.264 (0.193 to 0.335)	0.879 (0.864 to 0.893)	1.287 (1.193 to 1.380)	0.287 (0.193 to 0.380)
60-69 years	1,402 / 20,400	0.859 (0.847 to 0.870)	1.206 (1.140 to 1.273)	0.206 (0.140 to 0.273)	0.872 (0.856 to 0.888)	1.270 (1.166 to 1.374)	0.270 (0.166 to 0.374)
70-79 years	1,840 / 14,443	0.822 (0.811 to 0.833)	1.070 (1.014 to 1.127)	0.070 (0.014 to 0.127)	0.824 (0.807 to 0.841)	1.152 (1.059 to 1.245)	0.152 (0.059 to 0.245)
80+ years	3,158 / 11,915	0.762 (0.752 to 0.772)	0.874 (0.828 to 0.920)	-0.126 (-0.172 to -0.080)	0.740 (0.721 to 0.760)	0.834 (0.758 to 0.909)	-0.166 (-0.242 to -0.091)

Table 4.10. Age group-specific regression model performance metrics (with 95% confidence intervals) estimated after internal-external cross-validation in the data from period 2 (2010-2020). Harrell's C for the competing risks regression model was weighted by the inverse probability of censoring.

		XGBoost			Neural network		
Age group	Events / denominator	Harrell's C	Calibration slope	Calibration-in-the-large	Harrell's C*	Calibration slope	Calibration-in-the-large
20-29 years	26 / 317	0.787 (0.649 to 0.924)	0.914 (0.375 to 1.452)	-0.086 (-0.625 to 0.452)	0.840 (0.723 to 0.958)	0.881 (0.459 to 1.304)	-0.119 (-0.541 to 0.304)
30-39 years	287 / 3,259	0.714 (0.660 to 0.769)	1.316 (1.049 to 1.584)	0.316 (0.049 to 0.584)	0.769 (0.712 to 0.826)	1.017 (0.859 to 1.176)	0.017 (-0.141 to 0.176)
40-49 years	865 / 12,398	0.793 (0.766 to 0.820)	1.276 (1.140 to 1.412)	0.276 (0.140 to 0.412)	0.807 (0.780 to 0.833)	1.243 (1.136 to 1.350)	0.243 (0.136 to 0.350)
50-59 years	1,230 / 19,648	0.866 (0.850 to 0.882)	1.151 (1.074 to 1.228)	0.151 (0.074 to 0.228)	0.839 (0.820 to 0.859)	1.282 (1.209 to 1.356)	0.282 (0.209 to 0.356)
60-69 years	1,402 / 20,400	0.861 (0.845 to 0.876)	1.246 (1.153 to 1.338)	0.246 (0.153 to 0.338)	0.833 (0.814 to 0.851)	1.319 (1.233 to 1.405)	0.319 (0.233 to 0.405)
70-79 years	1,840 / 14,443	0.802 (0.787 to 0.818)	1.099 (1.005 to 1.194)	0.099 (0.005 to 0.194)	0.769 (0.749 to 0.788)	1.053 (0.986 to 1.120)	0.053 (-0.014 to 0.120)
80+ years	3,158 / 11,915	0.708 (0.687 to 0.730)	0.783 (0.719 to 0.847)	-0.217 (-0.281 to -0.153)	0.682 (0.659 to 0.704)	0.641 (0.595 to 0.687)	-0.359 (-0.405 to -0.314)

Table 4.11. Age group-specific machine learning model performance metrics (with 95% confidence intervals) estimated after internal-external cross-validation in the data from period 2 (2010-2020). Harrell's C for both models was weighted by the inverse probability of censoring.

		Cox proportional hazards model			Competing risks regression		
Stage at diagnosis	Events / denominator	Harrell's C	Calibration slope	Calibration-in-the-large	Harrell's C	Calibration slope	Calibration-in-the-large
Stage I	786 / 23,423	0.820 (0.804 to 0.834)	1.375 (1.285 to 1.465)	0.375 (0.285 to 0.465)	0.842 (0.818 to 0.866)	1.076 (1.013 to 1.138)	0.076 (0.013 to 0.138)
Stage II	2,067 / 21,473	0.776 (0.767 to 0.786)	1.191 (1.135 to 1.246)	0.191 (0.135 to 0.246)	0.796 (0.780 to 0.812)	1.314 (1.232 to 1.397)	0.314 (0.232 to 0.397)
Stage III	1,156 / 4,948	0.744 (0.731 to 0.757)	0.999 (0.929 to 1.069)	-0.001 (-0.071 to 0.069)	0.761 (0.736 to 0.786)	1.039 (0.833 to 1.244)	0.039 (-0.167 to 0.244)
Stage IV	1,708 / 3,011	0.713 (0.700 to 0.726)	0.803 (0.736 to 0.871)	-0.197 (-0.264 to -0.129)	0.681 (0.658 to 0.704)	0.837 (0.557 to 1.119)	-0.162 (-0.443 to 0.119)
		XGBoost			**Neural network**		
Stage I	786 / 23,423	0.810 (0.780 to 0.840)	1.535 (1.404 to 1.666)	0.535 (0.404 to 0.666)	0.756 (0.721 to 0.791)	1.684 (1.567 to 1.800)	0.684 (0.567 to 0.800)
Stage II	2,067 / 21,473	0.766 (0.747 to 0.786)	1.185 (1.121 to 1.249)	0.185 (0.121 to 0.249)	0.757 (0.735 to 0.780)	1.117 (1.066 to 1.168)	0.117 (0.066 to 0.168)
Stage III	1,156 / 4,948	0.678 (0.643 to 0.713)	0.614 (0.503 to 0.725)	-0.386 (-0.497 to -0.275)	0.676 (0.633 to 0.719)	0.602 (0.510 to 0.694)	-0.398 (-0.490 to -0.306)
Stage IV	1,708 / 3,011	0.620 (0.590 to 0.649)	0.276 (0.139 to 0.412)	-0.724 (-0.861 to -0.588)	0.621 (0.593 to 0.650)	0.126 (0.005 to 0.247)	-0.874 (-0.995 to -0.753)

Table 4.12. Tumour stage-specific summary performance metrics with 95% confidence intervals for the 4 models developed. Harrell's C was estimated using inverse probability of censoring weighting for the competing risks regression and machine learning models.

Group of predicted risk (highest)	Cox model		Competing risks regression		XGBoost		Neural network	
	Total breast cancer-related deaths in risk group	Cumulative % of total breast cancer-related deaths	Total breast cancer-related deaths in risk group	Cumulative % of total breast cancer-related deaths	Total breast cancer-related deaths in risk group	Cumulative % of total breast cancer-related deaths	Total breast cancer-related deaths in risk group	Cumulative % of total breast cancer-related deaths
1%	644	7.54%	631	7.16%	338	3.84%	610	6.93%
2%	1,244	14.12%	1,226	13.91%	837	9.50%	1,191	13.52%
3%	1,777	20.17%	1,766	20.04%	1,304	14.80%	1,715	19.47%
4%	2,257	25.62%	2,256	25.61%	1,729	19.63%	2,190	24.86%
5%	2,679	30.42%	2,728	30.97%	2,156	24.48%	2,655	30.14%
10%	4,306	48.89%	4,451	50.53%	3,765	42.75%	4,322	49.07%
15%	5,357	60.82%	5,489	62.31%	4,812	54.63%	5,255	59.66%
20%	6,120	69.48%	6,176	70.11%	5,564	63.17%	5,900	66.98%
25%	6,680	75.84%	6,701	76.08%	6,122	69.50%	6,427	72.97%
50%	8,124	92.23%	8,043	91.31%	7,626	86.58%	7,575	96.00%

Table 4.13. Sensitivity of each model for breast cancer death at different cut-offs of the predicted risks distribution.

The denominator is the number of breast cancer-related deaths occurring within Period 2 (n=8,808).

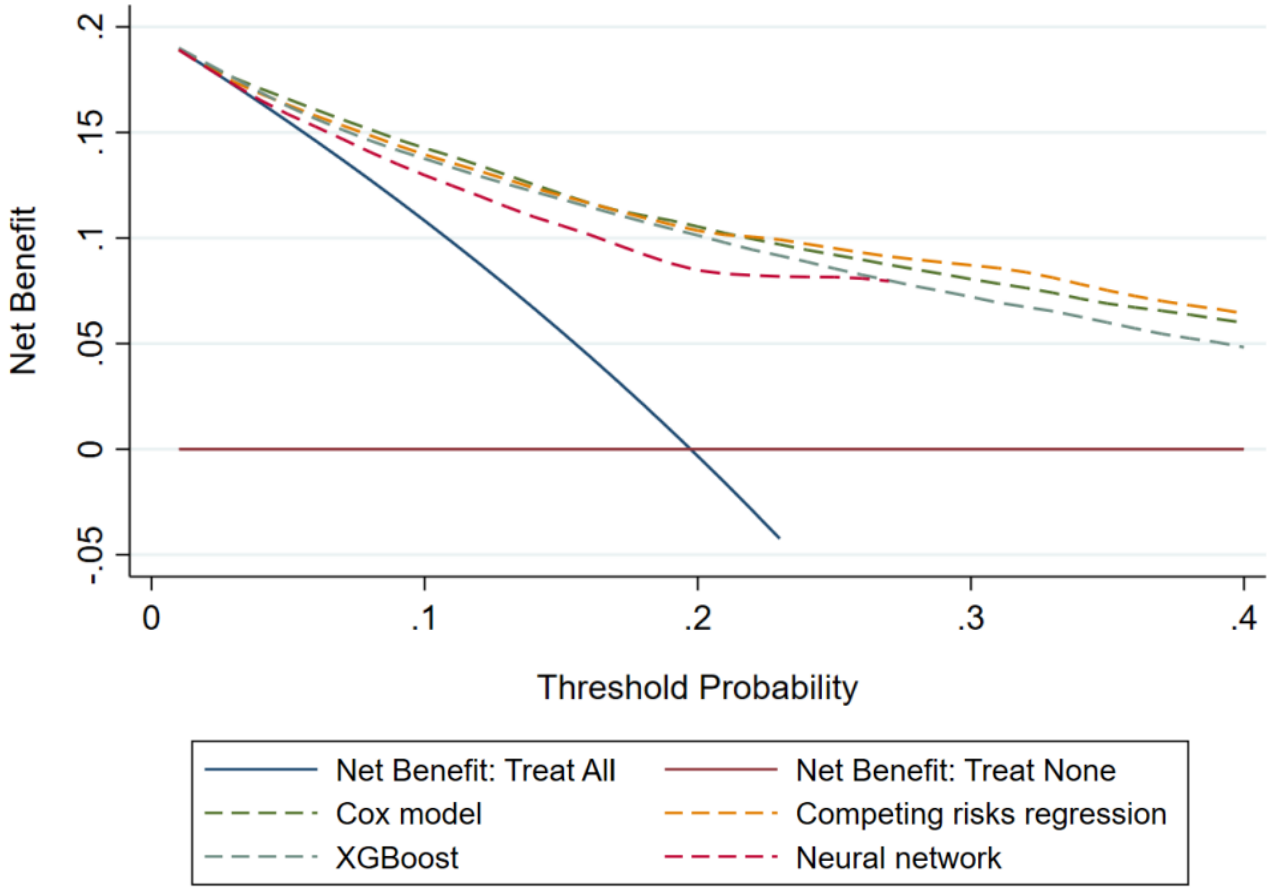

Figure 4.3. Decision curve analysis comparing the clinical utility (net benefit) of the 4 models developed, accounting for the competing risk of other-cause death.

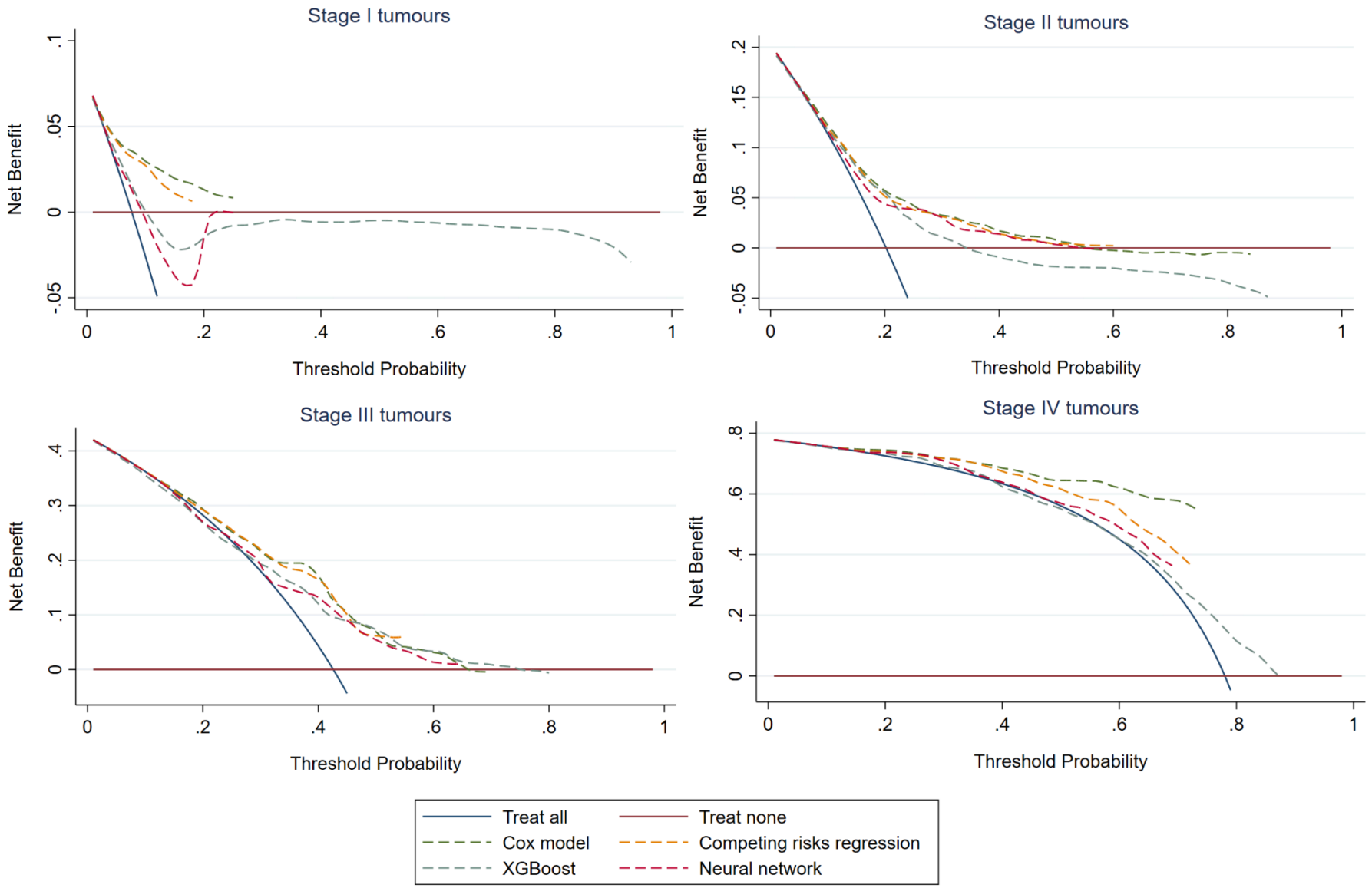

Figure 4.4. Tumour stage-specific decision curves comparing all 4 models.

		Cox model		Competing risks model		XGBoost		Neural network	
Performance metric	Variables in meta-regression model	I^2 (%)	R^2 (%)	I^2 (%)	R^2 (%)	I^2 (%)	R^2 (%)	I^2 (%)	R^2 (%)
Harrell's C-index	Age	43.06	0.00	41.45	9.54	17.64	0.00	43.68	8.05
	BMI	45.72	0.00	47.96	0.00	16.78	0.00	49.16	0.00
	Townsend deprivation score	44.28	0.00	46.25	0.00	15.87	3.34	48.29	0.00
	Non-white ethnicity	45.47	0.00	42.34	11.02	17.65	0.00	25.67	61.11
	All 4 variables	53.14	0.00	41.33	8.63	16.43	0.88	30.40	48.25
Calibration slope	Age	35.45	0.00	70.04	26.09	66.44	23.58	89.29	28.95
	BMI	35.93	0.00	77.14	0.00	74.72	0.00	93.20	0.00
	Townsend deprivation score	35.90	0.00	73.63	12.05	72.33	0.00	93.16	0.00
	Non-white ethnicity	34.21	0.00	58.49	55.77	56.06	51.91	87.07	42.31
	All 4 variables	42.35	0.00	56.68	60.08	50.18	61.97	67.78	80.06

Table 4.14. Random effects meta-regression to estimate relative contributions of regional variation in age, body mass index, deprivation and non-white ethnicity on inter-regional differences in performance metrics after internal-external cross-validation. Age and body mass index (BMI) are standard deviation, Townsend deprivation is mean, and 'non-white ethnicity' is the percentage (of those with recorded ethnicity) that were not of White ethnicity. I^2 = residual heterogeneity; R^2 = amount of residual heterogeneity accounted for.

Discussion

This chapter developed and comparatively evaluated four approaches to prognostication of women with breast cancer to identify the most promising models (as per the first research objective in **Chapter 1**). The regression approaches yielded models that discriminated well and were associated with favourable net benefit overall, but the XGBoost and neural network approaches yielded models that performed less uniformly. For example, the XGBoost and neural network models were associated with negative net benefit at some thresholds in Stage 1 tumours, were miscalibrated in stage III and IV tumours, and exhibited complex miscalibration across the spectrum of predicted risks.

Specific limitations of the work presented in this chapter include the inability to incorporate genetic or other 'omics' data that could have offered additional stratification utility to these models. Further, the approach taken did not enable causal prediction or consideration of treatment 'drop in' – models that are able to do so could theoretically be used to inform alterations to individual patient's lifestyle factors that could improve prognosis, for example, losing weight, or selecting treatments in counterfactual prediction scenarios[14,15]. The methods for doing so are not yet established and therefore the scope of this work was to explore methods to obtain accurate prognostication in a standard setting in the first instance.

This chapter along with **Chapters 3** and **5**, selected two machine learning approaches that could offer flexibility but could not be an exhaustive test of the panoply of tools that have been reported recently. The adaptations of XGBoost and neural networks used were developed to permit direct modelling of probabilities in a time-to-event scenario with relatively low-dimensional clinical data with an unknown signal-to-noise ratio, and could be repeatedly run within the IECV framework without being computationally

extortionate. Previous studies have adapted machine learning models using jack-knife pseudo-observations but these approaches do not appear to yet have wide adoption, nor have they been tested in terms of discrimination, calibration *and* clinical utility.

A recent model that was unable to be appraised by the Hueting, et al. systematic review due to its later publication date is Adjutorium[16]. Adjutorium is based on an ensemble machine learning approach named 'SurvivalQuilts', where multiple models are joined over timespans to estimate risks. The authors of the development paper concluded that this approach was superior to a Cox proportional hazards model in predicting risks in the post-surgical (adjuvant) setting. However this comparison is problematic as the statistical modelling failed to consider any interactions and only included a single, pre-determined (i.e. not data driven) non-linear term for age[16]. No modelling strategy represents a panacea in developing prognostic models using low dimension electronic healthcare data – appropriate, fair comparisons between different strategies are needed but not always achieved[17], and conclusions on best performing models should be based on broader criteria than single metrics. Further, on trialling the web-based calculator for this model, estimated risks of breast cancer mortality for screen-detected cancers were sometimes higher than for equivalent symptomatic cases, which may lead to counterintuitive/unexpected conclusions and affect face validity.

Although this chapter's intention was to develop models that are applicable to any woman diagnosed with breast cancer, a comparative evaluation of the PREDICT and Adjutorium models would have been an interesting analysis in the early breast cancer group treated with initial surgery. This was not possible due to some predictors required for these models being systematically missing in the available datasets.

As highlighted earlier, heterogeneity in model performance across clinically relevant sub-grouping is important to understand, especially in circumstances where thresholds may be low for initiating treatment[8,10]. This is inconsistently and rarely done. Measuring and understanding heterogeneity in model performance across societally relevant groups is also of interest if a model is intended to be deployed at a system level, due to concerns of algorithmic bias that could exacerbate existing healthcare inequities, or create new ones[18]. In this chapter, models were assessed in terms of their performance 'overall', by age group, ethnicity, and stage. Even the best performing models here were found to have clinically relevant variations in predictive performance in at least some groups. It is probably inappropriate to expect any model to 'work' equally well in all sub-groups, given that parameters are estimated using the average effects in a sample regardless of dataset size. The need for some recalibration or refitting in relevant groupings should be actively anticipated from the outset. As aforementioned, further study could include the updating of regression models with recalibration[19] or sub-group specific baseline functions, or fitting of stage-specific models using the machine learning approaches (as they do not contain a baseline term).

Regardless of the flexibility of modelling strategy used, all clinical prediction algorithms should be extensively evaluated and stress-tested. Showing a model 'works' overall should be subservient to understanding if, where and how a clinical prediction model will break down, and understanding what can be attempted to rectify performance in such scenarios. Subject to further evaluation, such as external validation and consideration of model updating to attain more uniform performance across stages, the Cox and competing risks regression models may have clinical utility worthy of future exploration.

This thesis now turns to developing and evaluating clinical prediction models for the final outcome in this work package – the 10-year risk of dying from breast cancer in women without breast cancer at baseline.

Chapter references

1. Siegfried S, Senn S, Hothorn T. On the relevance of prognostic information for clinical trials: A theoretical quantification. *Biom J* 2023; **65**(1): e2100349.

2. Wishart GC, Azzato EM, Greenberg DC, et al. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Res* 2010; **12**(1): R1.

3. Wishart GC, Bajdik CD, Dicks E, et al. PREDICT Plus: development and validation of a prognostic model for early breast cancer that includes HER2. *Br J Cancer* 2012; **107**(5): 800-7.

4. Haybittle JL, Blamey RW, Elston CW, et al. A prognostic index in primary breast cancer. *Br J Cancer* 1982; **45**(3): 361-6.

5. Todd JH, Dowle C, Williams MR, et al. Confirmation of a prognostic index in primary breast cancer. *Br J Cancer* 1987; **56**(4): 489-92.

6. Hayden JA, van der Windt DA, Cartwright JL, et al. Assessing bias in studies of prognostic factors. *Ann Intern Med* 2013; **158**(4): 280-6.

7. Phung MT, Tin Tin S, Elwood JM. Prognostic models for breast cancer: a systematic review. *BMC Cancer* 2019; **19**(1): 230.

8. Hueting TA, van Maaren MC, Hendriks MP, et al. The majority of 922 prediction models supporting breast cancer decision-making are at high risk of bias. *J Clin Epidemiol* 2022.

9. Engelhardt EG, van den Broek AJ, Linn SC, et al. Accuracy of the online prognostication tools PREDICT and Adjuvant! for early-stage breast cancer patients younger than 50 years. *Eur J Cancer* 2017; **78**: 37-44.

10. van Maaren MC, van Steenbeek CD, Pharoah PDP, et al. Validation of the online prediction tool PREDICT v. 2.0 in the Dutch breast cancer population. *Eur J Cancer* 2017; **86**: 364-72.

11. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Ann Intern Med* 2019; **170**(1): W1-W33.

12. Paredes-Aracil E, Palazon-Bru A, Folgado-de la Rosa DM, et al. Compan-Rosique AF, Gil-Guillen VF. A scoring system to predict breast cancer mortality at 5 and 10 years. *Sci Rep* 2017; **7**(1): 415.

13. Riley RD, Snell KI, 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**(7): 1276-96.

14. Lin L, Sperrin M, Jenkins DA, et al. A scoping review of causal methods enabling predictions under hypothetical interventions. *Diagn Progn Res* 2021; **5**(1): 3.

15. Sperrin M, Diaz-Ordaz K, Pajouheshnia R. Invited Commentary: Treatment Drop-in-Making the Case for Causal Prediction. *Am J Epidemiol* 2021; **190**(10): 2015-8.

16. Alaa AM, Gurdasani D, Harris AL, et al. Machine learning to guide the use of adjuvant therapies for breast cancer. *Nature Machine Intelligence* 2021; (3): 716-26.

17. Smith H, Sweeting M, Morris T, et al. A scoping methodological review of simulation studies comparing statistical and machine learning approaches to risk prediction for time-to-event data. *Diagn Progn Res* 2022; **6**(1): 10.

18. Rajpurkar P, Chen E, Banerjee O, et al. AI in health and medicine. *Nat Med* 2022; **28**(1): 31-8.

19. Booth S, Riley RD, Ensor J, et al. Temporal recalibration for improving prognostic model development and risk predictions in settings where survival is improving over time. *Int J Epidemiol* 2020; **49**(4): 1316-25.

Chapter Five

Prediction model development and evaluation – 10-year risk of breast cancer mortality in women without breast cancer at baseline

Summary

As summarised in **Chapter 1**, it should be considered whether structuring stratified screening or prevention strategies around individual estimations of incident breast cancer risks are optimal from the perspective of reducing breast cancer mortality. This chapter more deeply discusses the problems that breast cancer heterogeneity poses to such an approach, outlines the case for modelling an alternative, direct risk trajectory of breast cancer mortality, and then presents the results of this.

Introduction

A study using data collected from women in the Breast Cancer Surveillance Consortium (a screened population in the United States) calculated risk scores using three incident breast cancer models in women diagnosed with breast cancer, and found that predicted risks of incident cancer correlated poorly or even inversely with risk of mortality[1].

Evidence from the IBIS-I[2] and IBIS-II[3] trials demonstrated that whilst tamoxifen and anastrozole reduced breast cancer incidence by up to 35% by 15 years in women at increased risk (defined somewhat simplistically as having a relative risk of ≥2), the effects were heterogeneous across sub-types of the disease (including no effect in the prevention of ER- tumours, which have a worse prognosis than ER+ tumours) and there was no detectable effect on mortality. Lastly, screening may detect tumours that would never have otherwise threatened life or become clinically apparent (overdiagnosis), but nevertheless would be treated, presenting women with potential treatment risks for no benefit[4]. The extent of overdiagnosis is widely contested, and varies considerably depending on study approaches, definitions and assumptions[5-13].

Collectively, this supports the notion that merely seeking to identify any/all cancers may not necessarily be the most efficient or effective approach to reducing breast cancer mortality[4], due to the heterogeneity of risk trajectories of the disease itself. One must develop breast cancer in order to die from it, but cancers present differing threats to life depending on cancer type and patient characteristics. Increasing awareness of breast cancer heterogeneity has manifested in the risk modelling literature as efforts to evaluate model performance in predicting different disease sub-types[14], but also the development of models (clinical and/or polygenic in nature) to predict the risk of being diagnosed with tumours of these sub-types[15], or more recently, being diagnosed with advanced tumours (e.g. stage II and above)[16]. However, these models are typically built using data from women either within age criteria for screening or who have opted in to a screening programme[17-20], and models tend to perform less well for aggressive disease such as triple negative breast cancer[15].

This chapter takes a different approach by explicitly focussing on breast cancer mortality irrespective of women in the general population's eligibility for service screening. Being

able to identify women at increased risk of developing life-threatening cancer regardless of histological or receptor-defined type could offer an alternative approach to informing stratified early detection or chemoprevention.

Study population

The underlying cohort for this endpoint was identical to that used for modelling Endpoint One (**Chapter 3**) but with different endpoint and follow-up definitions, as per **Chapter 3**. After excluding women with breast cancer (GP/HES datasets, n=152,870) or ductal carcinoma *in situ* (n=5,409) diagnoses recorded prior to cohort entry, the final study cohort comprised the same 11,626,969 women included in the modelling in **Chapter 3**, as summarised in **Table 3.1**.

During a total of 70,095,574 person-years follow-up, there were 142,712 diagnoses of breast cancer (1.23%), 24,043 breast cancer-related deaths (0.21%), and 696,106 deaths due to other causes (5.99%). Median follow-up was 3.74 years (range 0.003 to 20.60), mean follow-up was 6.03 years. When restricting to 10-years of follow-up (prediction horizon), there were 13,062 breast cancer related deaths within 55,104,482 person years (crude mortality rate 2.37 per 10,000 person-years, 95% CI: 2.33 to 2.41).

In the temporally distinct sub-cohorts, there were 7,999 breast cancer deaths in Period 1 (crude mortality rate 2.66 per 10,000 person-years, 95% CI: 2.60 to 2.72), and 2,712 in Period 2 (1.54 per 10,000 person-years, 95% CI: 1.49 to 1.60). Crude ethnic group-specific and regional breast cancer mortality rates are summarised in **Tables 5.1** and **5.2**, respectively. Due to lower event counts for this outcome, a less granular categorisation of ethnicity was used.

Model development

Non-linear fractional polynomial terms were selected for age and BMI for the Cox model. Non-linearities were selected for age, BMI and Townsend deprivation score for the competing risks regression model (**Figures 5.1** and **5.2,** respectively). Interactions were considered between age and BMI, and age and family history of breast cancer.

Initially, ethnicity was selected for inclusion in both regression models as coefficients for several non-White ethnic groups were negative (i.e. associated hazard ratios <1). Whilst the predictor selection approach sought to identify variables significantly associated with the outcome of interest, it became apparent that this may have effects on model interpretation or potential later use in clinical decision making. As the envisioned use cases for these models include selecting women for interventions that seek to reduce mortality (e.g. screening or chemoprevention), using such models could lead to women in minority ethnic groups being systematically disadvantaged by the inclusion of this protected characteristic as a predictor. This presents an ethical issue that is discussed in more detail in the discussion section of this chapter. Therefore, the final models were developed excluding ethnicity as a candidate predictor.

Treatment-related variables were not included in these models as their values would not be known at the point of the prediction, i.e. in women that currently do not have breast cancer.

Although ethnicity was omitted as a predictor, performance heterogeneity remained of significant interest, and therefore the performance of both sets of models in different ethnic sub-groups was assessed. The final Cox and competing risks regression models (those not including ethnicity) are presented as their exponentiated coefficients with 95% confidence intervals in **Figures 5.3** and **5.4,** respectively, and reported in full in **Tables**

5.3 and **5.4**, respectively. The models including ethnicity as predictors are presented for the purpose of comparison in **Figures 5.5** and **5.6.**

Entire study cohort **Ethnic group**	**Number of person-years**	**Breast cancer deaths**	**Crude mortality rate per 10,000 person-years (95% confidence interval)**
White	32,333,543	5,154	1.59 (1.55 to 1.64)
South Asian	2,377,473.8	186	0.78 (0.68 to 0.90)
Other Asian	779,360.8	41	0.53 (0.39 to 0.71)
Black	1,975,478.2	243	1.23 (1.08 to 1.39)
Chinese	435,784.6	14	0.32 (0.19 to 0.54)
Other, inc. Arab and mixed race	1,203,062.1	69	0.57 (0.45 to 0.73)
Period 1 sub-cohort **Ethnic group**	**Number of person-years**	**Breast cancer deaths**	**Crude mortality rate per 10,000 person-years (95% confidence interval)**
White	16,106,040	1,734	1.08 (1.027 to 1.13)
South Asian	900,554.26	58	0.64 (0.50 to 0.83)
Other Asian	232,877.15	12	0.52 (0.29 to 0.91)
Black	752,790.48	72	0.96 (0.76 to 1.20)
Chinese	112,584.42	<5	0.36 (0.13 to 0.95)
Other, inc. Arab and mixed race	335,088.12	13	0.39 (0.23 to 0.67)
Period 2 sub-cohort **Ethnic group**	**Number of person-years**	**Breast cancer deaths**	**Crude mortality rate per 10,000 person-years (95% confidence interval)**
White	11,289,932	1,785	1.58 (1.51 to 1.66)
South Asian	1,027,905.8	70	0.68 (0.54 to 0.86)
Other Asian	385,655.87	12	0.31 (0.18 to 0.55)
Black	838,470.54	107	1.28 (1.06 to 1.54)
Chinese	249,231.92	5	0.20 (0.08 to 0.75)
Other, inc. Arab and mixed race	653,310.72	35	0.54 (0.38 to 0.75)

Table 5.1. Ethnic group-specific crude rates of breast cancer mortality in the study cohort overall (follow-up limited to 10-years from cohort entry to mimic the risk trajectory being modelled), and in the two temporally distinct sub-cohorts derived from splitting for internal-external validation. Due to the fact that some women entering during Period 1 may have contributed follow-up after 2010 in the 'overall cohort' and had an event therein, the temporal splitting and truncation of follow-up therein means that the numbers of events and rates in Period 1 plus Period 2 will not equate to those for the overall set.

Entire study cohort			
Region	**Number of person-years**	**Breast cancer deaths**	**Crude mortality rate per 10,000 person-years (95% confidence interval)**
East Midlands	2,455,094.7	748	3.05 (2.84 to 3.27)
East of England	3,165,611.7	862	2.72 (2.54 to 2.91)
London	13,310,057	1,938	1.46 (1.39 to 1.52)
North East	1,819,810.7	590	3.24 (2.99 to 3.51)
North West	8,983,125.3	2,357	2.62 (2.52 to 2.73)
South Central	6,567,956.7	1,548	2.36 (2.42 to 2.48)
South East	4,737,608.4	1,250	2.64 (2.50 to 2.79)
South West	5,795,710.6	1,467	2.53 (2.40 to 2.66)
West Midlands	5,530,304.9	1,569	2.84 (2.70 to 2.98)
Yorkshire & Humber	2,739,201.6	733	2.68 (2.49 to 2.88)
Total	*55,104,482*	*13,062*	*2.37 (2.33 to 2.41)*
Period 1 sub-cohort			
Region	**Number of person-years**	**Breast cancer deaths**	**Crude mortality rate per 10,000 person-years (95% confidence interval)**
East Midlands	1,668,925.1	490	2.94 (2.69 to 3.21)
East of England	1,985,918.2	539	2.71 (2.49 to 2.95)
London	6,284,964.3	1,118	1.78 (1.68 to 1.89)
North East	1,167,361.1	320	2.74 (2.46 to 3.05)
North West	4,743,542.2	1,415	2.98 (2.83 to 3.14)
South Central	3,708,440.6	1,006	2.71 (2.55 to 2.89)
South East	2,382,666.8	719	3.02 (2.80 to 3.25)
South West	3,291,332.3	906	2.75 (2.58 to 2.94)
West Midlands	3,081,618	968	3.14 (2.95 to 3.35)
Yorkshire & Humber	1,783,093.7	518	2.91 (2.67 to 3.17)
Total	*30,097,862*	*7,999*	*2.66 (2.60 to 2.72)*
Period 2 sub-cohort			
Region	**Number of person-years**	**Breast cancer deaths**	**Crude mortality rate per 10,000 person-years (95% confidence interval)**
East Midlands	471,558	138	2.92 (2.48 to 3.46)
East of England	748,256.02	143	1.91 (1.62 to 2.25)
London	5,106,112.3	458	0.90 (0.82 to 0.98)
North East	311,775.65	77	2.47 (1.98 to 3.09)

North West	2,975,473	505	1.70 (1.56 to 1.85)
South Central	2,010,833.1	307	1.53 (1.37 to 1.71)
South East	1,733,116.9	292	1.68 (1.50 to 1.90)
South West	1,830,815.7	341	1.86 (1.67 to 2.07)
West Midlands	1,747,196.3	349	2.00 (1.80 to 2.22)
Yorkshire & Humber	573,631.02	94	1.64 (1.34 to 2.01)
Total	*17,508,768*	*2,704*	*1.54 (1.49 to 1.60)*

Table 5.2. Regional crude rates of breast cancer mortality in the study cohort overall (follow-up limited to 10-years from cohort entry to mimic the risk trajectory being modelled), and in the two temporally distinct sub-cohorts derived from splitting for internal-external validation. Due to the fact that some women entering during Period 1 may have contributed follow-up after 2010 in the 'overall cohort' and had an event therein, the temporal splitting and truncation of follow-up therein means that the numbers of events and rates in Period 1 plus Period 2 will not equate to those for the overall set.

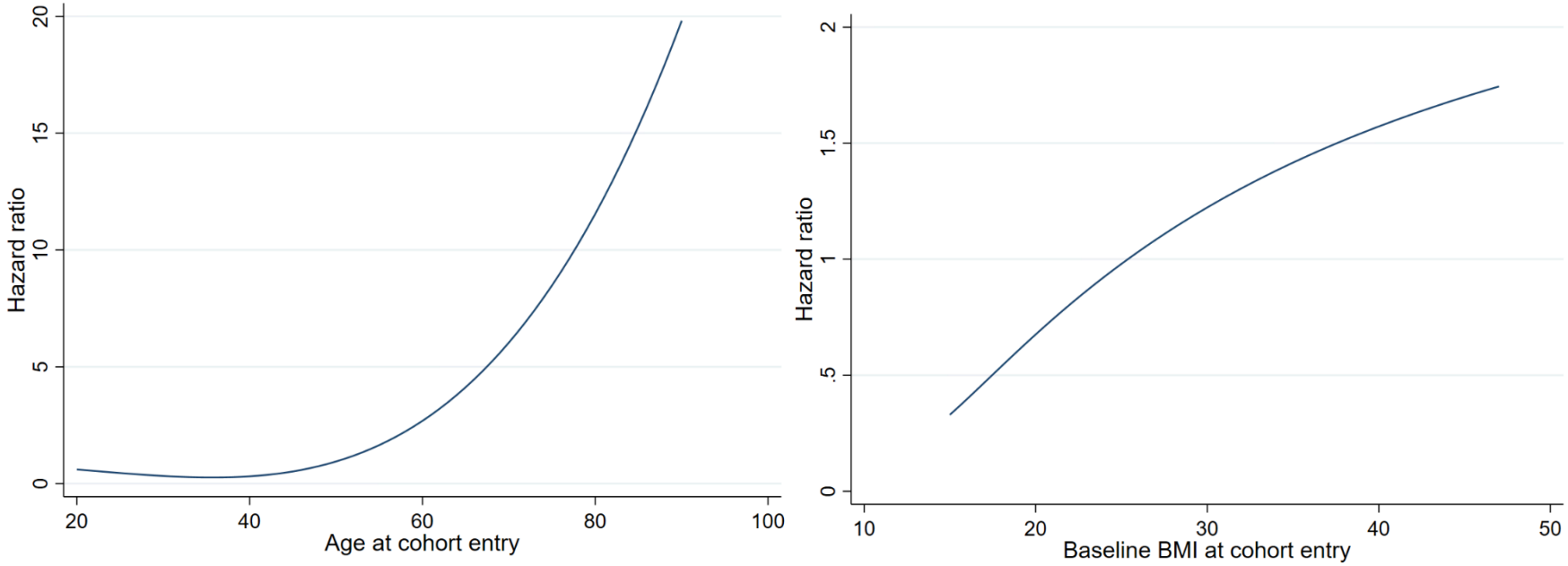

Figure 5.1. Fractional polynomial terms selected for age (left, [-2,3]) and body mass index (right, [-2,-2]) for the Cox proportional hazards model.

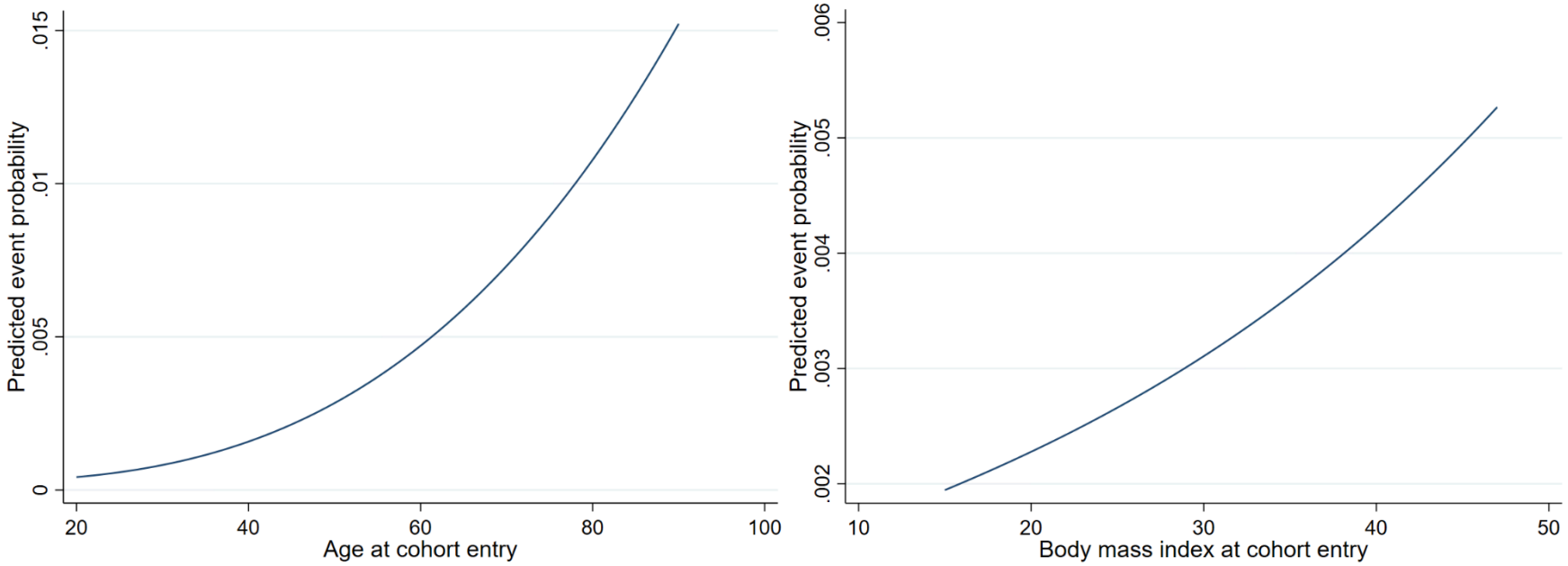

Figure 5.2. Fractional polynomial terms selected for age (left [1,2]) and body mass index (right, [1]) for the competing risks regression model.

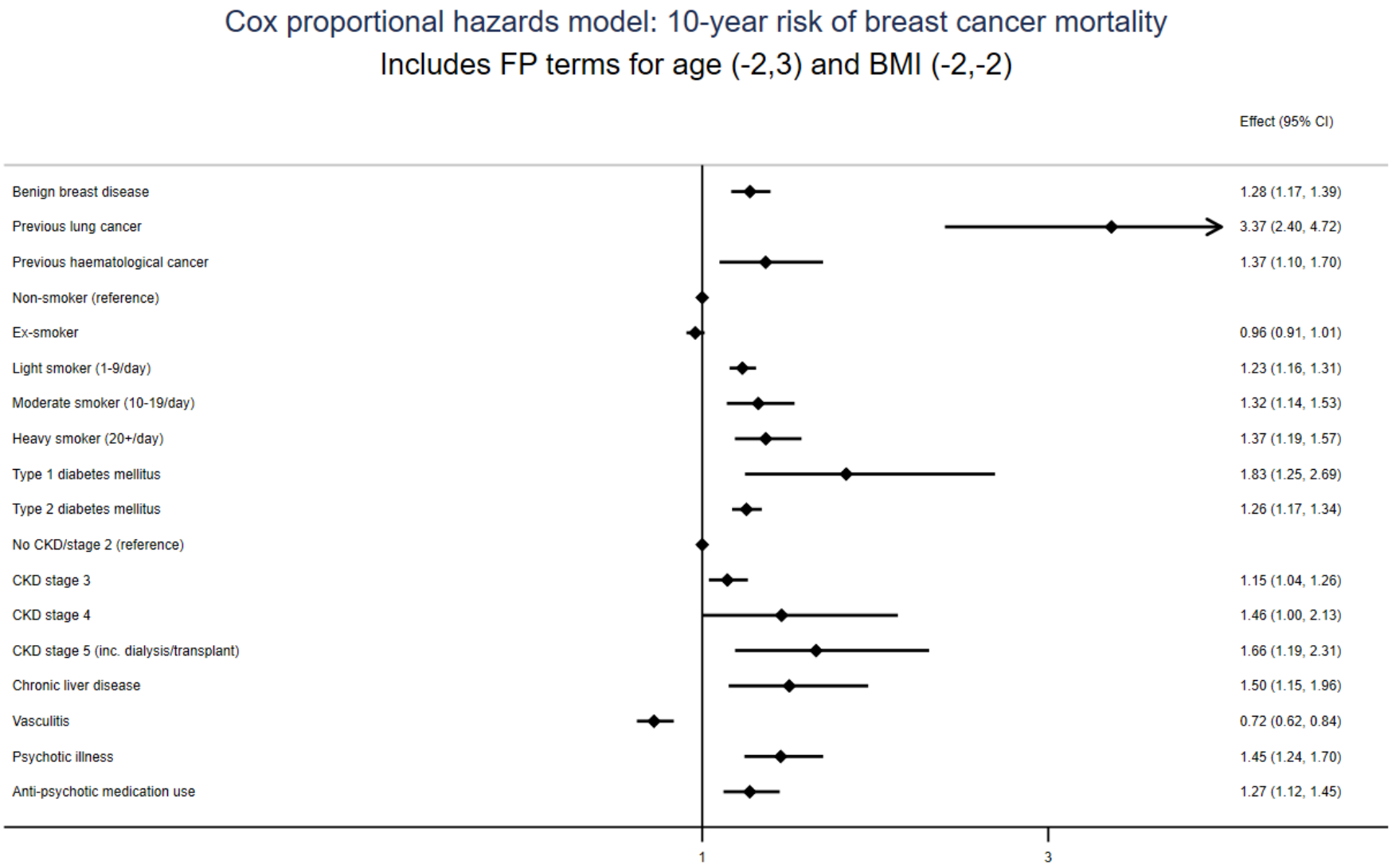

Figure 5.3. Final Cox proportional hazards model presented as exponentiated coefficients to graphically summarise the relative effects of individual parameters on risk estimates. BMI = body mass index; CKD = chronic kidney disease; CI = confidence interval.

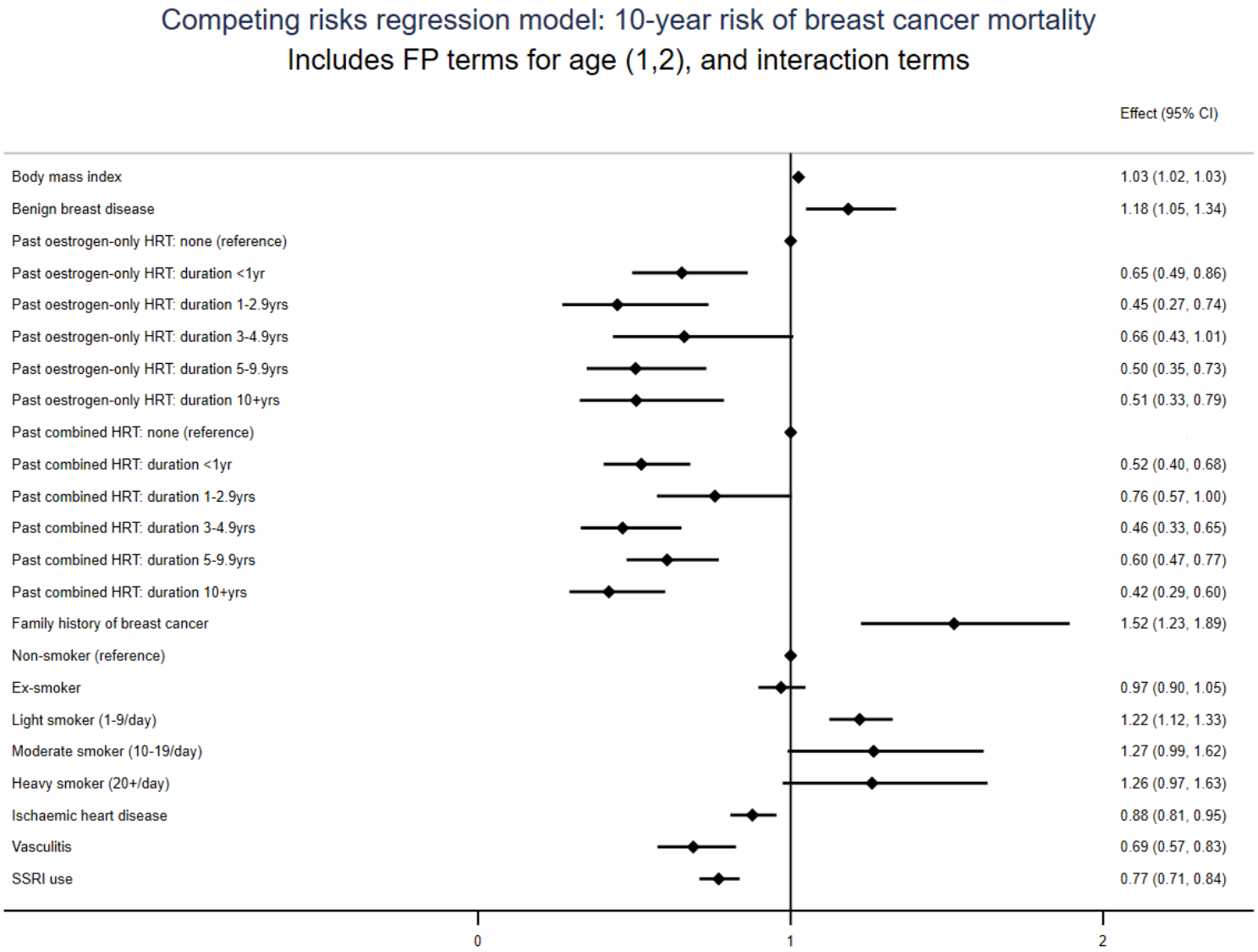

Figure 5.4. Final competing risks regression model presented as exponentiated coefficients to graphically summarise the relative effects of individual parameters on risk estimates. HRT = hormone replacement therapy, SSRI = selective serotonin reuptake inhibitor.

Parameter	Category/description	Coefficient
Age at entry – FP term 1	X^{-2} – 0.55476233 *X = age/10*	0.09702326
Age at entry– FP term 2	X^{3} – 76.53138187 *X = age/10*	-0.00432757
BMI – FP term 1	X^{-2} – 0.15531752 *X = BMI/10*	1.5420898
BMI – FP term 2	X^{-2} * ln(X) – 0.1446226461 *X = BMI/10*	-7.2562036
Benign breast disease	No (reference) Yes	0 0.24399999
Previous lung cancer	No (reference) Yes	0 1.2135882
Previous haematological cancer	No (reference) Yes	0 0.31245335
Smoking status	Non-smoker (reference) Ex-smoker Light smoker (1-9/day) Moderate smoker (10-19/day) Heavy smoker (20+/day)	0 -0.04098904 0.20879024 0.28043591 0.31306449
Type 1 diabetes mellitus	No (reference) Yes	0 0.60552787
Type 2 diabetes mellitus	No (reference) Yes	0 0.22760716
Chronic kidney disease	None/stage 2 (reference) Stage 3 Stage 4 Stage 5/ESRF/dialysis/transplant	0 0.13567373 0.37760089 0.5056187
Chronic liver disease	No (reference) Yes	0 0.40758873
Vasculitis	No (reference) Yes	0 -0.32661846
History of psychotic condition	No (reference) Yes	0 0.37422545
Anti-psychotic medication use *(3+ prescriptions ever)*	No (reference) Yes	0 0.24284637
Baseline survival function at 10 years		0.9998084

Table 5.3. Final Cox proportional hazards model coefficients and baseline survival function. FP = fractional polynomial, BMI = body mass index (kg/m^2), ESRF = end-stage renal failure.

Parameter	Category/description	Coefficient
Age at entry – FP term 1	X – 4.24567277 X = age/10	1.0099973
Age at entry – FP term 2	X^2 – 18.02573727 X = age/10	-0.04055561
BMI	(linear)	0.0248805
Past use of oestrogen-only HRT *(last prescription 5+ years ago)*	None (reference) <1 year duration 1 - 2.9 years duration 3 – 4.9 years duration 5 – 9.9 years duration 10+ years duration	0 -0.42811746 -0.80836627 -0.41595355 -0.68563903 -0.68148758
Past use of combined HRT *(last prescription 5+ years ago)*	None (reference) <1 year duration 1 - 2.9 years duration 3 – 4.9 years duration 5 – 9.9 years duration 10+ years duration	0 -0.64970264 -0.27761333 -0.77160656 -0.50290198 -0.8714718
Family history of breast cancer	No (reference) Yes	0 0.42068759
Smoking status	Non-smoker (reference) Ex-smoker Light smoker (1-9/day) Moderate smoker (10-19/day) Heavy smoker (20+/day)	0 -0.03117927 0.19969482 0.23555489 0.23148372
History of ischaemic heart disease	No (reference) Yes	0 -0.13037047
Vasculitis	No (reference) Yes	0 -0.37358314
SSRI use *(3+ prescriptions ever)*	No (reference) Yes	0 -0.2617108
Interaction: age (FP term 1) * family history of breast cancer		-1.0437374
Interaction: age (FP term 2) * family history of breast cancer		0.07335707
Constant term		-7.0619127

Table 5.4. Final competing risks regression model coefficients and constant term. FP = fractional polynomial, BMI = body mass index (kg/m2), HRT = hormone replacement therapy, SSRI = selective serotonin reuptake inhibitor use.

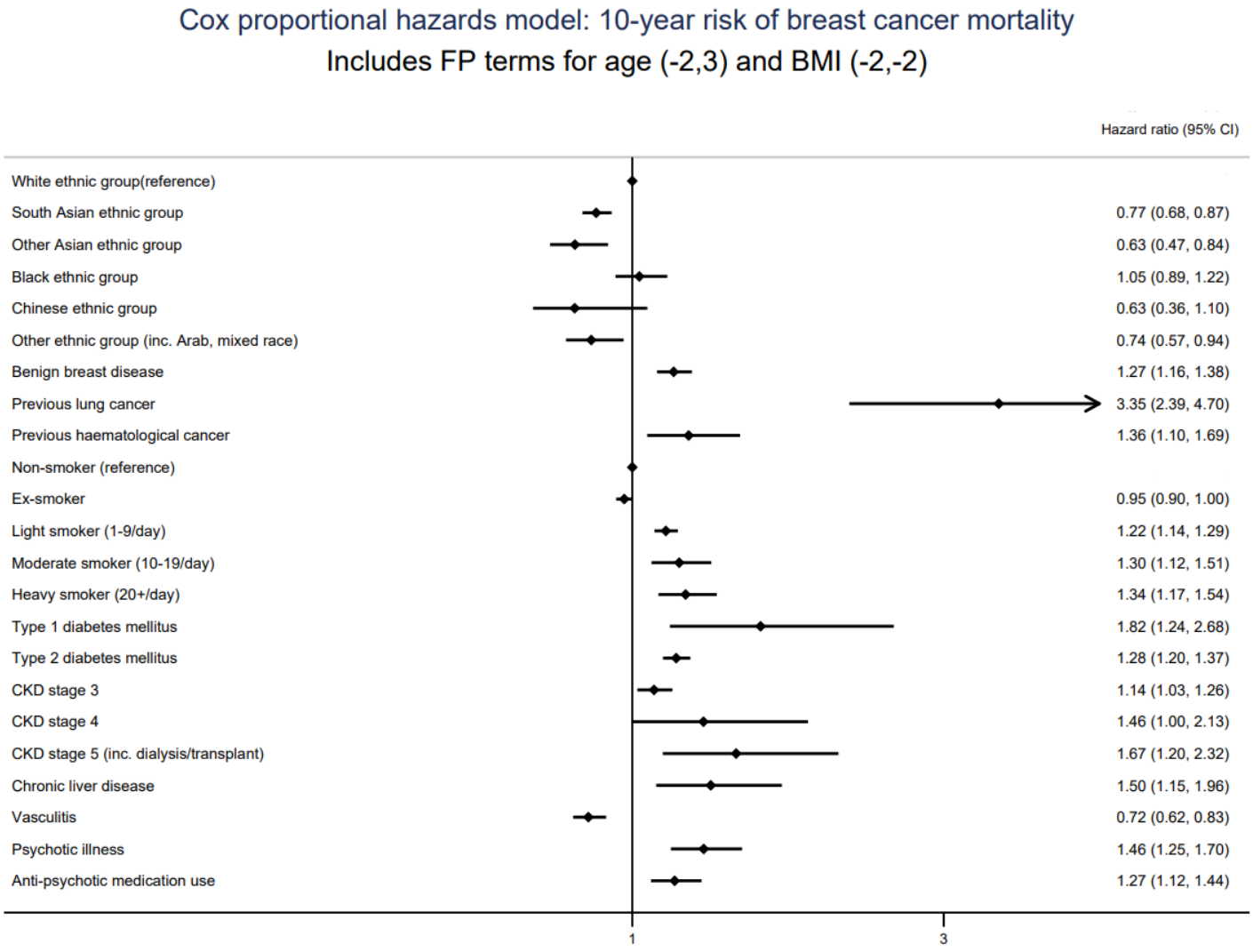

Figure 5.5. Cox proportional hazards model including ethnicity – displayed as exponentiated coefficients (hazard ratios) and confidence intervals. BMI = body mass index; CKD = chronic kidney disease.

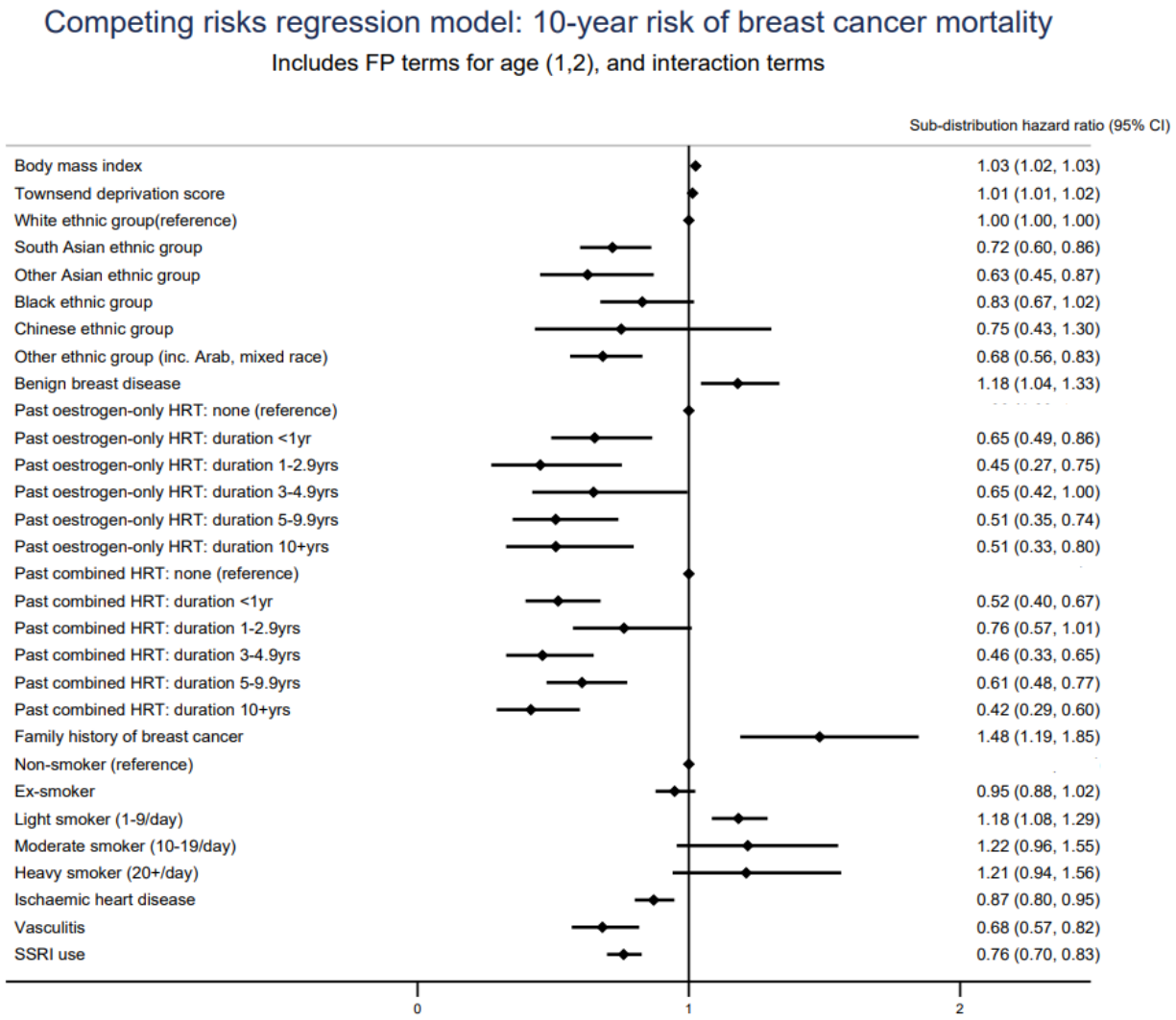

Figure 5.6. Competing risks model including ethnicity – displayed as exponentiated coefficients (sub-distribution hazards ratios) and confidence intervals.

Model	Basic architecture	Hyperparameters tuned	Range explored during hyperparameter tuning	Final selected value after tuning
XGBoost	Tree-based booster with 'GPU_hist' method	Maximum tree depth	1 to 6	4
		Learning rate (eta)	0 to 0.1	0.065
		Sub-sample proportion	0.1 to 0.8	0.236
	Gradient-based sub-sampling	Number of boosting rounds	0 to 500	448
		Alpha (regularisation)	0 to 20	16
	'Reg:squared error' objective	Gamma (regularisation)	0 to 20	0
		Lambda (regularisation)	0 to 20	9
	RMSE evaluation metric	Column sampling by tree	0.1 to 0.8	0.650
		Column sampling by level	0.1 to 0.8	0.258
Neural network	Feed-forward ANN with fully connected layers	Number of hidden layers	1 to 5	4
		Number of nodes in each hidden layer	24 to 48	48
	ReLU activation functions in hidden layers	Number of epochs	1 to 10	3
		Initial learning rate	0.001 to 0.1	0.001
	Adam optimiser			
	Single output node with linear activation			
	RMSE loss function			
	Batch size 8192			

Table 5.5. Hyperparameter tuning, and final configurations of the machine learning models predicting 10-year risk of breast cancer mortality. RMSE = root mean squared error, ANN = artificial neural network. ReLU = rectified linear unit. The batch size was based on consideration of run time (model fit, and internal-external validation on a dataset with over 50million observations within a maximum of 7 days with graphical processing unit support). A sensitivity analysis using a batch size of 1024 had lower Harrell's C statistic (0.655). XGBoost used 50 iterations of Bayesian Optimization, the neural network used 20 iterations due to the smaller hyperparameter search space and computational intensity of model fitting.

Table 5.5 summarises the hyperparameter tuning search spaces and final selected configurations for the XGBoost and neural network models. The final neural network had a total of 8,305 parameters. Following interim analysis of results from the regression models, the machine learning modelling was undertaken in a competing risks framework.

Model evaluation – overall performance

Table 5.6 summarises the performance metrics estimated using internal-external cross-validation for all 4 models predicting the 10-year risk of breast cancer mortality.

The competing risks regression model displayed the highest discrimination with a Harrell's C statistic of 0.931 (95% CI: 0.917 to 0.946, 95% prediction interval: 0.886 to 0.977), whilst the neural network had the lowest (Harrell's C statistic 0.771, 95% CI: 0.751 to 0.792, 95% prediction interval: 0.718 to 0.792).

The Cox, competing risks and XGBoost models did not display any discernible miscalibration on summary measures. The neural network did, however, although this was to a negligible extent, with a calibration slope of 1.037 (95% CI: 1.003 to 1.071, 95% prediction interval 0.935 to 1.140).

On visualisation of calibration plots (**Figure 5.7**), all methods tended towards overestimation at the very highest range of the predicted risk spectrum; miscalibration of the Cox model began at a lower range, suggesting variations in transportability across models.

All models tended towards overestimation in those with the very highest predicted risks (e.g. >2% 10-year risk). The degree to which such miscalibration may be clinically important should extend beyond calibration curve shape – for the competing risks model for example, the tendency towards overestimation appeared to occur above a risk threshold of 0.015. This represented only 0.7% of patients (denominator was the number of individuals in the Period 2 sub-cohort), but perhaps more salient to consider is the importance of miscalibration (i.e. overestimation) in a group that has at least a 1.5% risk of developing and then dying from breast cancer within 10-years, which is likely to be deemed a high-risk group in many population settings.

Model	Harrell's C index	Calibration slope	Calibration in the large
Cox model	0.854 (0.842 to 0.865) [0.822 to 0.885]	1.091 (0.991 to 1.191) [0.787 to 1.395]	0.091 (-0.009 to 0.191) [-0.213 to 0.395]
Competing risks regression	0.931 (0.917 to 0.946) [0.885 to 0.978]	1.011 (0.978 to 1.044) [0.913 to 1.110]	0.011 (-0.022 to 0.044) [-0.087 to 0.110]
XGBoost	0.839 (0.805 to 0.873) [0.737 to 0.942]	1.021 (0.989 to 1.052) [0.926 to 1.116]	0.021 (-0.011 to 0.052) [-0.074 to 0.116]
Neural network	0.771 (0.751 to 0.792) [0.718 to 0.792]	1.037 (1.003 to 1.071) [0.935 to 1.140]	0.037 (0.003 to 0.071) [-0.065 to 0.140]
Model	**Royston & Sauerbrei's D Statistic**	**Royston & Sauerbrei's R^2**	**Brier score**
Cox model	2.397 (2.288 to 2.506) [2.117 to 2.677]	0.579 (0.557 to 0.601) [0.523 to 0.636]	0.003 (0.002 to 0.003) [0.001 to 0.004]

Table 5.6. Summary performance metrics for the 4 models developed to predict 10-year risk of breast cancer mortality. Royston & Sauerbrei's D Statistic and R^2 are not estimable for the competing risks (regression or machine learning) models. Harrell's C at 10 years was weighted by inverse probability of censoring for the competing risks (regression and machine learning) models.

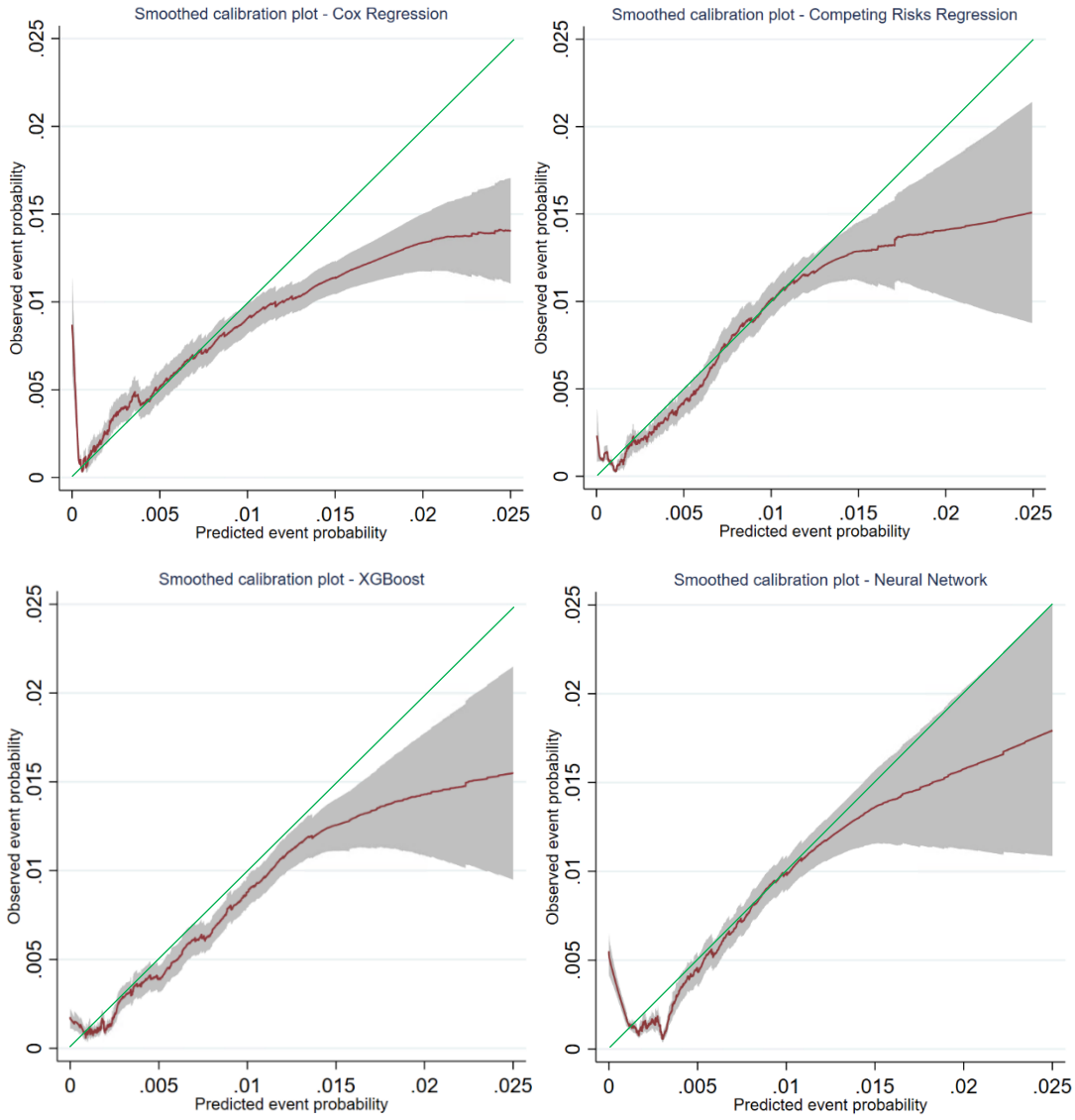

Figure 5.7. Smoothed calibration plots for each model predicting 10-year risk of breast cancer mortality. Observed probabilities were estimated using pseudo-observations for the Kaplan-Meier failure function (Cox) or Aalen-Johansen cumulative incidence function (other) at 10 years.

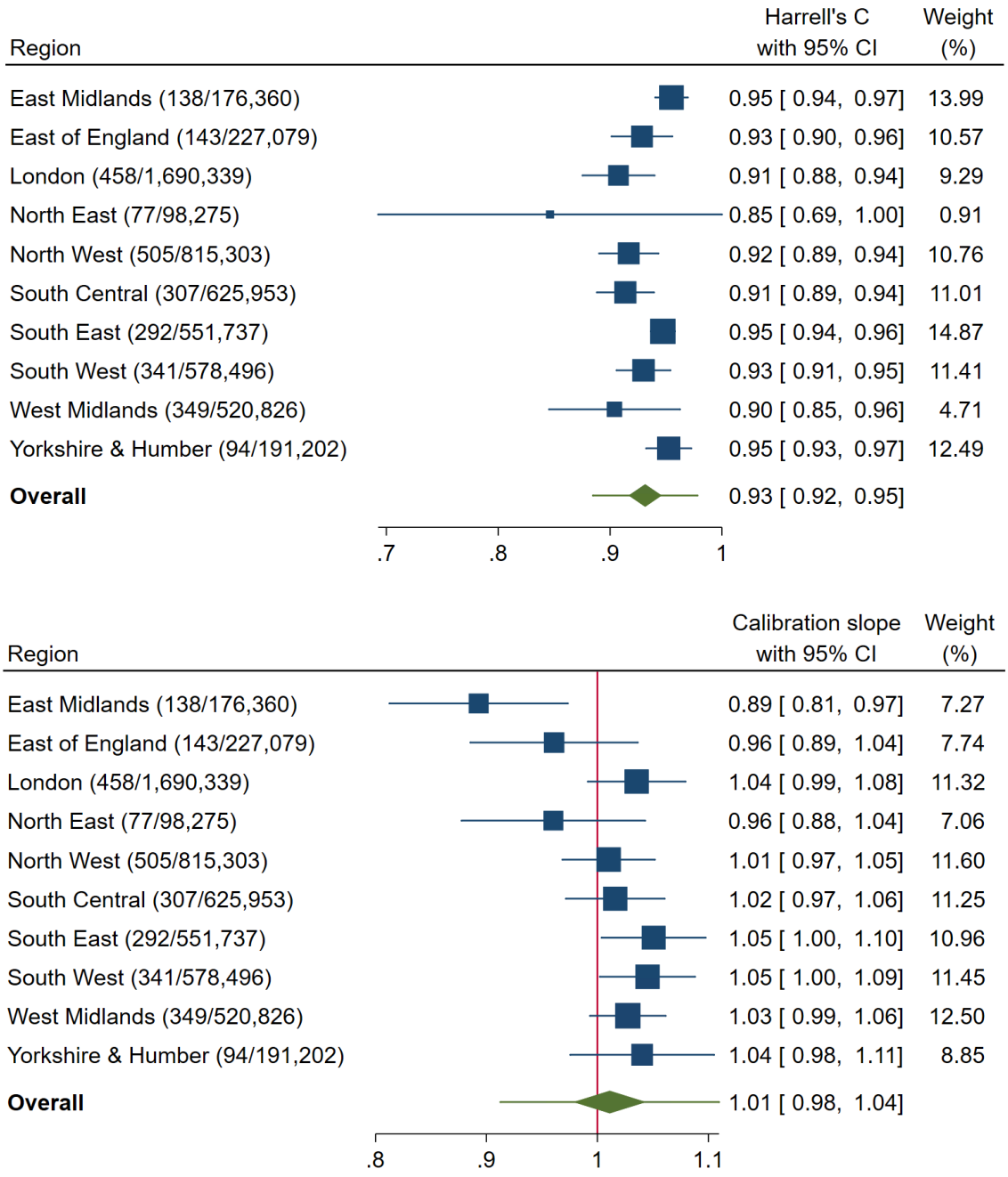

Figure 5.8. Performance metrics (top – Harrell's C; bottom – calibration slope) for the competing risks regression model obtained using random effects meta-analysis pooling after internal-external cross-validation. The green diamond on each forest plot represents the meta-estimate with the 95% confidence interval, and the line spanning outwards therefrom represents the 95% prediction interval. The red line on the calibration slope plot refers to the 'ideal' value of 1.

Model evaluation – performance heterogeneity

Ethnic group performance

The final regression models generally appeared to have acceptable discrimination across ethnic sub-groups (**Table 5.7**), although the confidence intervals for the 'Other Asian' sub-group were wide, reflecting low event counts even in this large sample with ensuing reduced precision.

With the caveat of small numbers of events, these models were generally well calibrated in most ethnic groups, apart from negligible miscalibration with the Cox model in White women (slope 1.060, 95% CI: 1.016 to 1.105), and with the competing risks model in the 'Other Asian' sub-group (slope 1.252, 95% CI: 1.075 to 1.428).

Interestingly, the XGBoost and neural network approaches had less stable performance across different ethnic groups (**Table 5.8**). This manifested, for example, as poor discrimination of both models in Black women (Harrell's C statistic for XGBoost: 0.569, 95% CI: 0.418 to 0.720; for neural network: 0.623, 95% CI: 0.469 to 0.776).

Age group performance

As the envisioned use case of these models includes identifying women that are higher risk but may be 'unrecognised' and fall outside the current screening eligibility criteria, it was of interest to estimate how well each model performed in women of different age ranges. Three sub-groups informed by age-based screening eligibility of the NHS Breast

Screening Programme were explored: 'pre-screening' women aged 20-49 years, 'screening-age' women aged 50-70years, and 'post-screening' women aged 71 years and above. Of note, due to the cohort eligibility criteria, these post-screening women could have partaken in screening in earlier life but were never diagnosed with breast cancer or DCIS.

When exploring age-based sub-group performance, more complex patterns of performance were observed (**Table 5.9**). For example, the Cox model generally discriminated better than the XGBoost and neural network models, but it was miscalibrated in younger women (slope 1.771, 95% CI: 1.558 to 1.954) and those older than current screening age (slope 0.120, 95% CI: -0.108 to 0.349). Discrimination of the machine learning models in younger women was poor (e.g. Harrell's C for XGBoost 0.404, 95% CI: 0.359 to 0.449). The competing risks regression model did not show any miscalibration in any screening-relative age group – discrimination was lowest in women of screening age, but higher in the pre- and post-screening age sub-groups.

		Cox proportional hazards model			Competing risks regression model		
Ethnic group	Events / denominator	Harrell's C	Calibration slope	Calibration-in-the-large	Harrell's C*	Calibration slope	Calibration-in-the-large
White	1,785 / 3,417,048	0.858 (0.850 to 0.865)	1.060 (1.016 to 1.105)	0.060 (0.0156 to 0.105)	0.923 (0.909 to 0.937)	1.013 (0.995 to 1.031)	0.013 (-0.005 to 0.031)
South Asian	70 / 307,926	0.851 (0.802 to 0.900)	1.161 (0.984 to 1.338)	0.161 (-0.016 to 0.338)	0.885 (0.826 to 0.944)	1.041 (0.945 to 1.138)	0.041 (-0.055 to 0.138)
Other Asian	12 / 123,097	0.748 (0.541 to 0.954)	0.981 (0.352 to 1.611)	-0.018 (-0.648 to 0.611)	0.730 (0.502 to 0.959)	1.252 (1.075 to 1.428)	0.252 (0.075 to 0.428)
Black	107 / 255,632	0.782 (0.731 to 0.834)	0.903 (0.745 to 1.061)	-0.097 (-0.255 to 0.061)	0.747 (0.638 to 0.857)	1.025 (0.929 to 1.122)	0.025 (-0.071 to 0.122)
Chinese	5 / 92,940	0.877 (0.737 to 1.000)	1.168 (0.471 to 1.866)	0.168 (-0.529 to 0.866)	0.904 (0.755 to 1.000)	1.130 (0.994 to 1.265)	0.130 (-0.006 to 0.265)
Other ethnic group, including Arab and mixed race	35 / 227,249	0.867 (0.818 to 0.916)	1.013 (0.768 to 1.258)	0.013 (-0.232 to 0.258)	0.907 (0.850 to 0.964)	1.064 (0.954 to 1.173)	0.064 (-0.046 to 0.173)

Table 5.7. Ethnic group-specific summary performance metrics for all the regression models predicting 10-year risk of breast cancer mortality. These are estimated using internal-external validation, and therefore counts correspond to events in period 2 data. * = estimated using inverse probability of censoring weighting, due to competing risks formulation.

		XGBoost			Neural network		
Ethnic group	Events / denominator	Harrell's C*	Calibration slope	Calibration-in-the-large	Harrell's C*	Calibration slope	Calibration-in-the-large
White	1,785 / 3,417,048	0.863 (0.847 to 0.880)	1.023 (1.005 to 1.040)	0.023 (0.005 to 0.040)	0.788 (0.767 to 0.809)	1.040 (1.022 to 1.058)	0.040 (0.022 to 0.058)
South Asian	70 / 307,926	0.697 (0.572 to 0.822)	1.072 (0.981 to 1.163)	0.072 (-0.019 to 0.163)	0.828 (0.716 to 0.940)	1.081 (0.991 to 1.171)	0.081 (-0.008 to 0.171)
Other Asian	12 / 123,097	0.486 (0.060 to 0.911)	1.190 (1.049 to 1.331)	0.190 (0.049 to 0.331)	0.694 (0.390 to 0.998)	1.204 (1.075 to 1.334)	0.204 (0.075 to 0.334)
Black	107 / 255,632	0.569 (0.418 to 0.720)	1.042 (0.956 to 1.129)	0.042 (-0.044 to 0.129)	0.623 (0.469 to 0.776)	1.060 (0.993 to 1.127)	0.060 (-0.007 to 0.127)
Chinese	5 / 92,940	0.653 (0.282 to 1.000)	1.100 (1.028 to 1.172)	0.100 (0.028 to 0.172)	0.851 (0.685 to 1.00)	1.135 (1.074 to 1.196)	0.134 (0.074 to 0.196)
Other ethnic group, including Arab and mixed race	35 / 227,249	0.723 (0.574 to 0.871)	1.064 (0.962 to 1.166)	0.064 (-0.038 to 0.166)	0.750 (0.608 to 0.892)	1.090 (0.898 to 1.191)	0.089 (-0.011 to 0.191)

Table 5.8. Ethnic group-specific summary performance metrics for the machine learning models predicting 10-year risk of breast cancer mortality. These are estimated using internal-external validation, and therefore counts correspond to events in period 2 data. * = estimated using inverse probability of censoring weighting, due to competing risks formulation.

		Cox proportional hazards model			Competing risks regression model		
Age group	Events / denominator	Harrell's C	Calibration slope	Calibration-in-the-large	Harrell's C*	Calibration slope	Calibration-in-the-large
20-49 years	502 / 3,992,447	0.760 (0.738 to 0.783)	1.771 (1.588 to 1.954)	0.771 (0.587 to 0.954)	0.735 (0.682 to 0.788)	1.024 (0.994 to 1.053)	0.024 (-0.006 to 0.053)
50-70 years	822 / 1,009,935	0.594 (0.570 to 0.617)	0.705 (0.554 to 0.857)	-0.295 (-0.446 to -0.143)	0.579 (0.553 to 0.605)	1.027 (0.991 to 1.063)	0.027 (-0.009 to 0.063)
>70 years	1,380 / 473,188	0.619 (0.601 to 0.636)	0.120 (-0.108 to 0.349)	-0.880 (-1.108 to -0.651)	0.636 (0.618 to 0.654)	1.012 (0.991 to 1.033)	0.012 (-0.009 to 0.033)
		XGBoost			**Neural network**		
Age group	Events / denominator	Harrell's C*	Calibration slope	Calibration-in-the-large	Harrell's C*	Calibration slope	Calibration-in-the-large
20-49 years	502 / 3,992,447	0.404 (0.359 to 0.449)	1.056 (1.034 to 1.079)	0.056 (0.034 to 0.079)	0.522 (0.474 to 0.570)	1.136 (1.108 to 1.163)	0.136 (0.108 to 0.163)
50-70 years	822 / 1,009,935	0.586 (0.561 to 0.611)	1.016 (0.979 to 1.052)	0.016 (-0.021 to 0.052)	0.547 (0.521 to 0.572)	1.049 (1.010 to 1.088)	0.049 (0.010 to 0.088)
>70 years	1,380 / 473,188	0.587 (0.566 to 0.608)	1.026 (1.004 to 1.048)	0.026 (0.004 to 0.048)	0.578 (0.556 to 0.601)	1.000 (0.976 to 1.02)	0.000 (-0.024 to 0.023)

Table 5.9. Age group-specific summary performance metrics for all 4 models. These are estimated using internal-external validation, and therefore counts correspond to events in period 2 data. * = estimated using inverse probability of censoring weighting, due to competing risks formulation.

Clinical utility

Sensitivity for breast cancer deaths

The sensitivity of each model was assessed by calculating the proportion of breast cancer deaths accounted for in different parts of their respective predicted risk distributions. 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 (**Table 5.10**). The competing risks model captured 54% of all breast cancer deaths in the top 10% of predicted risks.

Net benefit – decision curve analysis

Decision curve analysis was performed overall, and separately for each age-based sub-group (**Figure 5.9**).

Overall, the neural network was associated with the lowest net benefit, whereas the other models were associated with similar-to-or-better net benefit compared to the 'screen/treat all' strategy. Of note, the 'screen/treat all' strategy is unrealistic as a public health strategy due to the logistical, clinical and economic burdens of screening all women aged 20 years and above, as well as the uncertainties regarding efficacy of screening outside of middle age for the general population.

The regression models were associated with improved net benefit in women below screening age as per the NHSBSP (**Figure 5.9**), and the competing risks model could be

the most useful for clinical decision-making in women older than screening age (**Figure 5.9**). The latter is perhaps not wholly unexpected, as it is in this age group that the competing risks of death from other causes are most prominent.

All models except the neural network were associated with modest difference in net benefit in the screening-age group compared to 'screen all' strategy, suggesting that using these other models to inform strategy in screening-age women may be at least as good as screening all women within the remit of that age group. This, however, should be interpreted in the context of performance heterogeneity across ethnic groups.

Group of predicted risk (highest)	Cox model		Competing risks regression		XGBoost		Neural network	
	Total breast cancer-related deaths in risk group	Cumulative % of total breast cancer-related deaths	Total breast cancer-related deaths in risk group	Cumulative % of total breast cancer-related deaths	Total breast cancer-related deaths in risk group	Cumulative % of total breast cancer-related deaths	Total breast cancer-related deaths in risk group	Cumulative % of total breast cancer-related deaths
1%	244	9.02%	265	9.80%	232	8.58%	251	9.28%
2%	431	15.94%	491	18.16%	453	16.75%	453	16.75%
3%	620	22.93%	666	24.63%	644	23.82%	602	22.26%
4%	788	29.14%	846	31.29%	808	29.88%	765	28.29%
5%	927	34.28%	994	36.76%	952	35.21%	892	32.99%
10%	1,450	53.62%	1,481	54.77%	1,468	54.29%	1,349	49.89%
15%	1,734	64.13%	1,746	64.57%	1,756	64.94%	1,600	59.17%
20%	1,972	72.93%	1,955	72.30%	1,953	72.23%	1,766	65.31%
25%	2,147	79.40%	2,120	78.40%	2,106	77.88%	1,846	68.27%
50%	2,570	95.04%	2,561	94.71%	2,317	85.69%	2,181	80.66%

Table 5.9. Proportion of breast cancer-related deaths captured by varying groups of predicted risks, for each of the 4 models.

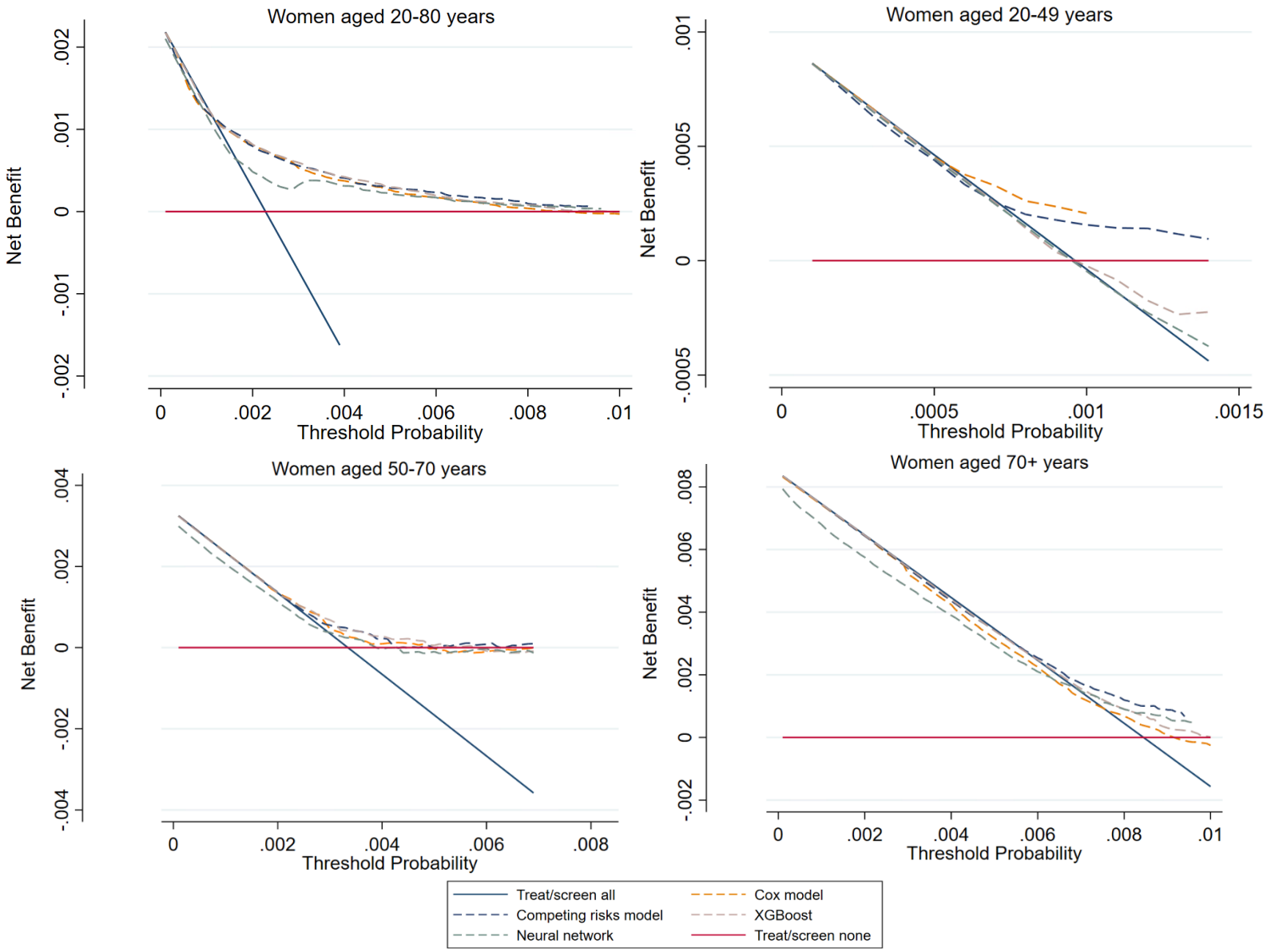

Figure 5.9. Decision analysis curves for (top left) the Period 2 sub-cohort, (top right) women below screening eligibility, (bottom left) women eligible for screening, and (bottom right) women older than screening eligibility.

Performance of models including ethnicity as a predictor

The Cox proportional hazards model including ethnicity as a predictor had a higher discrimination point estimate than the model that did not include it but overlapping confidence and prediction intervals – Harrell's C estimated after IECV 0.885 (95% CI: 0.842 to 0.867, 95% prediction interval 0.821 to 0.888) versus 0.854 (95% CI: 0.854 to 0.865, 95% prediction interval 0.822 to 0.885). The effect on the competing risks model of not including ethnicity was a negligible decrement in Harrell's C of 0.006, and the competing risks regression model remained the one with highest discrimination when comparing models that included ethnicity – Harrell's C 0.937 (95% CI: 0.919 to 0.954, 95% prediction interval 0.883 to 0.990).

Regarding calibration, including ethnicity had a minimal effect on calibration (e.g. slope 1.011, 95% CI: 0.978 to 1.044) for the competing risks model. However, it induced a tiny degree of miscalibration for the Cox model – slope 1.103 (95% CI: 1.099 to 1.185) compared to a slope of 1.091 (95% CI: 0.991 to 1.191) for the 'race-blind' Cox model.

Inclusion of ethnicity as a predictor was associated with minimal effects on the summary discrimination and calibration metrics across ethnic groups (**Tables 5.10 & 5.11**) – point estimates were similar and confidence intervals were overlapping.

Meta-regression

Random effects meta-regression estimated the relative contributions of inter-region variation in the heterogeneity of results for the summary performance metrics. The results are summarised in **Table 5.12**.

		Cox proportional hazards model			Competing risks regression model		
Ethnic group	Events / denominator	Harrell's C	Calibration slope	Calibration-in-the-large	Harrell's C*	Calibration slope	Calibration-in-the-large
White	1,785 / 3,417,048	0.858 (0.851 to 0.865)	1.070 (1.021 to 1.118)	0.070 (0.021 to 0.118)	0.922 (0.902 to 0.942)	1.015 (0.996 to 1.033)	0.015 (-0.004 to 0.033)
South Asian	70 / 307,926	0.849 (0.798 to 0.900)	1.165 (0.988 to 1.342)	0.165 (-0.012 to 0.342)	0.891 (0.828 to 0.954)	1.008 (0.909 to 1.107)	0.008 (-0.091 to 0.107)
Other Asian	12 / 123,097	0.768 (0.570 to 0.966)	1.007 (0.389 to 1.625)	0.007 (-0.612 to 0.625)	0.745 (0.538 to 0.953)	1.197 (1.003 to 1.391)	0.197 (0.003 to 0.391)
Black	107 / 255,632	0.782 (0.729 to 0.835)	0.911 (0.754 to 1.068)	-0.089 (-0.256 to 0.068)	0.759 (0.642 to 0.876)	1.056 (0.947 to 1.165)	0.056 (-0.053 to 0.165)
Chinese	5 / 92,940	0.876 (0.735 to 1.000)	1.173 (0.487 to 1.860)	0.173 (-0.513 to 0.859)	0.891 (0.724 to 1.000)	1.135 (0.858 to 1.411)	0.135 (-0.142 to 0.412)
Other ethnic group, including Arab and mixed race	35 / 227,249	0.861 (0.808 to 0.914)	0.990 (0.742 to 1.238)	-0.010 (-0.257 to 0.238)	0.922 (0.872 to 0.973)	1.038 (0.930 to 1.147)	0.038 (-0.070 to 0.147)

Table 5.10. Summary performance metrics for the regression models predicting 10-year risk of breast cancer mortality, but also including ethnicity as a predictor.

		XGBoost			Neural network		
Ethnic group	Events / denominator	Harrell's C	Calibration slope	Calibration-in-the-large	Harrell's C*	Calibration slope	Calibration-in-the-large
White	1,785 / 3,417,048	0.850 (0.830 to 0.870)	1.049 (1.029 to 1.069)	0.049 (0.029 to 0.069)	0.788 (0.767 to 0.809)	1.040 (1.022 to 1.058)	0.040 (0.022 to 0.058)
South Asian	70 / 307,926	0.688 (0.558 to 0.818)	1.122 (1.027 to 1.218)	0.122 (0.027 to 0.218)	0.828 (0.716 to 0.940)	1.081 (0.991 to 1.171)	0.081 (-0.008 to 0.171)
Other Asian	12 / 123,097	0.488 (0.000 to 1.000)	1.194 (1.058 to 1.331)	0.194 (0.058 to 0.331)	0.694 (0.390 to 0.998)	1.204 (1.075 to 1.334)	0.204 (0.075 to 0.334)
Black	107 / 255,632	0.649 (0.532 to 0.765)	1.063 (0.948 to 1.178)	0.063 (-0.052 to 0.177)	0.623 (0.469 to 0.776)	1.060 (0.993 to 1.127)	0.060 (-0.007 to 0.127)
Chinese	5 / 92,940	0.536 (0.128 to 0.944)	1.154 (1.098 to 1.209)	0.154 (0.098 to 0.209)	0.851 (0.685 to 1.00)	1.135 (1.074 to 1.196)	0.134 (0.074 to 0.196)
Other ethnic group, including Arab and mixed race	35 / 227,249	0.703 (0.538 to 0.871)	1.091 (0.997 to 1.185)	0.091 (-0.003 to 0.185)	0.750 (0.608 to 0.892)	1.090 (0.898 to 1.191)	0.089 (-0.011 to 0.191)

Table 5.11. Summary performance metrics for the machine learning models predicting 10-year risk of breast cancer mortality, but also including ethnicity as a predictor.

		Cox model		Competing risks model		XGBoost		Neural network	
Performance metric	Variables in meta-regression model	I^2 (%)	R^2 (%)	I^2 (%)	R^2 (%)	I^2 (%)	R^2 (%)	I^2 (%)	R^2 (%)
Harrell's C	Age	33.24	54.67	72.21	3.52	63.08	40.55	20.47	29.42
	BMI	52.10	2.94	74.19	0.00	69.72	11.51	15.42	49.42
	Townsend score	56.86	0.00	68.14	5.44	74.39	0.00	29.30	0.00
	Non-white ethnicity	44.78	28.05	69.13	17.83	75.44	0.00	22.96	19.79
	All 4 variables	40.05	35.47	65.97	3.85	47.90	57.13	11.74	60.05
Calibration slope and calibration-in-the-large	Age	78.19	7.39	72.94	0.00	72.05	0.00	73.51	0.00
	BMI	79.87	3.55	72.29	0.00	72.16	0.00	72.53	0.00
	Townsend score	80.13	10.30	71.38	0.47	70.43	1.76	72.60	0.55
	Non-white ethnicity	81.47	0.00	69.85	5.81	68.15	9.99	71.42	0.87
	All 4 variables	74.22	25.86	70.00	1.44	70.08	0.00	62.21	32.20

Table 5.12. Random effects meta-regression to estimate relative contributions of regional variation in age, body mass index, deprivation and non-white ethnicity on inter-regional differences in performance metrics after internal-external validation. Age and body mass index (BMI) are standard deviation, Townsend deprivation is mean, and 'non-white ethnicity' is the percentage (of those with recorded ethnicity data) that were not of white ethnicity. Estimates of I^2 and R^2 were derived from univariate, and then multivariate meta-regression models. I^2 = residual heterogeneity; R^2 = amount of residual heterogeneity accounted for.

Discussion

The competing risks regression model was deemed the best performing model. It had the highest discrimination of any model, was generally well calibrated, and performed acceptably both overall and across selected age- and ethnic- sub-groups.

The strengths and limitations of the work undertaken in this Chapter mirror those discussed in **Chapter 3,** which developed and evaluated models to predict breast cancer incidence. Pertinent to this chapter is the consideration to which it is limited by the inability to incorporate genetic risk estimates or mammographic density, due to non-availability in the source datasets which comprise routinely collected electronic healthcare data. Whilst incorporating such parameters has yielded incremental benefit in terms of increased discrimination indices in models predicting future risks of incident breast cancer diagnoses[21-23], their relevance and potential impact on this chapter's risk trajectory is not certain and requires further, specific evaluation.

For example, one group has reported that efforts to develop polygenic risk scores (models that aggregate the effects of individual genetic variants on risk) for breast cancer death using the UK Biobank yielded an unstable model that was not recommendable for use[24]. Whilst breast density is a recognised to be associated with breast cancer incidence/diagnosis risks, it has no association with breast cancer mortality[25]. Further, as the envisioned use cases for the models in this chapter include identification of unrecognised high-risk women that may be 'missed' by current age-based screening eligibility, higher-risk women that are too young to be screened are unlikely to have recent mammography imaging available; offering a 'one-off' screen for the whole adult female

population to provide data for population stratification is likely to be logistically challenging and medico-ethically questionable.

An important theme that emerged in this chapter was the ethics of including protected characteristics such as ethnicity in clinical prediction models which could influence clinical decision making or eligibility to healthcare services. Self-reported ethnicity was initially selected for inclusion in both regression models due to the magnitudes of association on risk of breast cancer mortality in women without breast cancer at baseline. Causal interpretations of these model coefficients cannot be applied, but all other factors being equal, non-white women would receive systematically lower risk estimates than women of White backgrounds with identical predictor values if these models were implemented. In the context of an intended use case being identification of women for mortality-reducing interventions, use of the ethnicity-including models could manifest as reduced access to stratified early detection or prevention services, thereby predicating or exacerbating health inequities despite improving outcomes at the overall, whole-population level.

Whilst ethnic groups differ in terms of their crude breast cancer incidence risks (Cancer Research UK data[26], and **Table 3.5**), these observations may be attributable to different 'risk factor' distributions therein[28,29]. Further, after a diagnosis of breast cancer has been made, survival outcomes may vary across ethnicities (**Table 4.2**)[28,29]. This chapter models a combined risk trajectory with two 'steps' – developing a cancer that is diagnosed, and then dying from it. Predictors could be associated with either of these components or both; if associated with both, they may be so with different directions or magnitude of association.

Being able to 'explain' the risk estimate from the ethnicity-including model clinical prediction model output to a woman from a minority ethnic background would not only be complex from the perspective of the process within the calculation mechanism, but could also raise concerns regarding trust. A recent citizen's jury explored views on the use of the QCovid prediction model in informing public health policy for COVID-19[27], but the conclusions are transportable to other scenarios. One of the 'red lines' of this consensus approach was the use of a risk tool that leads to service access denial based on ethnicity. Therefore, implementing an ethnicity-including model from this chapter may have reduced face validity and acceptability with intended end-beneficiaries.

Crucially, model coefficients for ethnic groups may not reflect true, immutable biological characteristics. Rather, they are probably a proxy reflecting structural /socio-contextual factors alongside any biological effect, if one exists[28,29]. Excluding this predictor could represent a form of omitted variable bias – omitting ethnicity from the Cox proportional hazards model was associated with a small decrement to the discrimination capability of that model, but it had negligible effect on the competing regression model which remained the highest performing. Considering all of these factors, the thesis presented the 'ethnicity-blind' models as its primary models.

Accurate tools that can identify women at increased risk of developing life-threatening breast cancers could inform efficient targeting of those most likely to benefit from chemoprevention, novel screening approaches, or recruitment into trials studying these interventions. They could also synergise with efforts directed towards reducing breast cancer incidence pending further evaluation, such as randomised impact assessments, external local validation, and health economic evaluations.

This thesis now turns to an assessment of the potential incremental effects of integrating additional predictors (not available in the source data) into this model and the best performing models for Endpoint One (breast cancer incidence). As per the second research objective outlined in **Chapter 1**, it of interest to compare the performance of models that could function using routinely collected electronic healthcare record data only, versus those that integrate data fields that are either not well recorded or non-routine in primary care.

Chapter references

1. 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**(8): e0160966.
2. 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**(1): 67-75.
3. 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**(10218): 117-22.
4. Autier P, Boniol M. Mammography screening: A major issue in medicine. *Eur J Cancer* 2018; **90**: 34-62.

5. Blyuss O, Dibden A, Massat NJ, et al. A case-control study to evaluate the impact of the breast screening programme on breast cancer incidence in England. *Cancer Med* 2022.

6. Cox B, Sneyd MJ. Bias in breast cancer research in the screening era. *Breast* 2013; **22**(6): 1041-5.

7. Etzioni R, Gulati R. Recognizing the Limitations of Cancer Overdiagnosis Studies: A First Step Towards Overcoming Them. *J Natl Cancer Inst* 2016; **108**(3).

8. Jorgensen KJ, Gotzsche PC, Kalager M, et al. Breast Cancer Screening in Denmark: A Cohort Study of Tumor Size and Overdiagnosis. *Ann Intern Med* 2017; **166**(5): 313-23.

9. Kalager M, Adami HO, Bretthauer M, et al. Overdiagnosis of invasive breast cancer due to mammography screening: results from the Norwegian screening program. *Ann Intern Med* 2012; **156**(7): 491-9.

10. Lee CI, Etzioni R. Missteps in Current Estimates of Cancer Overdiagnosis. *Acad Radiol* 2017; **24**(2): 226-9.

11. Ryser MD, Etzioni RB. Estimation of Breast Cancer Overdiagnosis in a U.S. Breast Screening Cohort. *Ann Intern Med* 2022; **175**(10): W116-W7.

12. Ryser MD, Gulati R, Eisenberg MC, et al. Identification of the Fraction of Indolent Tumors and Associated Overdiagnosis in Breast Cancer Screening Trials. *Am J Epidemiol* 2019; **188**(1): 197-205.

13. Welch HG, Prorok PC, O'Malley AJ, et al. Breast-Cancer Tumor Size, Overdiagnosis, and Mammography Screening Effectiveness. *N Engl J Med* 2016; **375**(15): 1438-47.

14. 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**(5): 489-97.

15. 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**(1): 21-34.

16. Kerlikowske K, Chen S, Golmakani MK, et al. Cumulative Advanced Breast Cancer Risk Prediction Model Developed in a Screening Mammography Population. *J Natl Cancer Inst* 2022; **114**(5): 676-85.

17. Shieh Y, Hu D, Ma L, et al. Breast cancer risk prediction using a clinical risk model and polygenic risk score. *Breast Cancer Res Treat* 2016; **159**(3): 513-25.

18. Shieh Y, Hu D, Ma L, et al. Joint relative risks for estrogen receptor-positive breast cancer from a clinical model, polygenic risk score, and sex hormones. *Breast Cancer Res Treat* 2017; **166**(2): 603-12.

19. Tice JA, Bissell MCS, Miglioretti DL, et al. Validation of the breast cancer surveillance consortium model of breast cancer risk. *Breast Cancer Res Treat* 2019; **175**(2): 519-23.

20. Tice JA, Cummings SR, Ziv E, et al. Mammographic breast density and the Gail model for breast cancer risk prediction in a screening population. *Breast Cancer Res Treat* 2005; **94**(2): 115-22.

21. Brentnall AR, Cuzick J, Buist DSM, et al. Long-term Accuracy of Breast Cancer Risk Assessment Combining Classic Risk Factors and Breast Density. *JAMA Oncol* 2018; **4**(9): e180174.

22. Evans DGR, van Veen EM, Harkness EF, et al. Breast cancer risk stratification in women of screening age: Incremental effects of adding mammographic density, polygenic risk, and a gene panel. *Genet Med* 2022; **24**(7): 1485-94.

23. 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**(4): 476-82.

24. Neale Lab. SNP Heritability for Phenotype 40001_C509. 2022. https://nealelab.github.io/UKBB_ldsc/h2_summary_40001_C509.html.

25. Gierach GL, Ichikawa L, Kerlikowske K, et al. Relationship between mammographic density and breast cancer death in the Breast Cancer Surveillance Consortium. *J Natl Cancer Inst* 2012; **104**(16): 1218-27.

26. Cancer Research UK. Breast cancer statistics. 2022. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer.

27. IpsosScotland & Scottish Government. Citizen's Jury on QCovid: Report on the jury's conclusions and key findings. 2022. https://www.gov.scot/publications/citizens-jury-qcovid-report-jurys-conclusions-key-findings/#:~:text=QCovid%C2%AE%20is%20a%20risk,Citizens'%20Jury%20process%20and%20findings.

28. Gathani T, Ali R, Balkwill A, et al. Ethnic differences in breast cancer incidence in England are due to differences in known risk factors for the disease: prospective study. *Br J Cancer* 2014; **110**(1): 224-9.

29. Gathani T, Chaudhry A, Chagla L, et al. Ethnicity and breast cancer in the UK: Where are we now? *Eur J Surg Oncol* 2021; **47**(12): 2978-81.

Chapter Six

Methods for external evaluation and integrated prediction modelling in UK Biobank

Summary

Pursuant to developing and evaluating models for 10-year risk of incident breast cancer diagnosis and 10-year risk of breast cancer mortality, this thesis now considers their performance in an external dataset, and the effects on performance of integrating additional data types that were not available in QResearch. Using the UK Biobank, a cohort study approach is used to externally evaluate the Cox proportional hazards model for Endpoint 1, and the competing risks regression model for Endpoint 2. Thereafter, an existing genome-wide polygenic score is combined with the linear predictors of these phenotypic models, and also reproductive factors, which are not available or are poorly recorded in the QResearch derivation dataset.

Introduction

Evaluation of clinical prediction models in populations independent to the one(s) used for model development can be useful to provide additional information on generalisability of the model to new settings[1]. Non-random splitting within an internal-external validation framework in very large data can provide impressions of external validity[2], but independent populations may have dissimilar baseline risks (e.g. the incidence rate of the outcome of interest is different), or variations in predictor distributions or predictor-outcome associations, any of which may affect model performance[1,2]. External validation datasets are also samples of their underlying population, so biases pertaining to selection should also be considered.

The UK Biobank is a deeply phenotyped and genotyped prospective cohort study[3,4]. Over 500,000 adults aged 40-70 years accepted an invitation to participate (approximately 6% of all invited). The UK Biobank has collected data from individuals attending a baseline assessment centre to record anthropometrics and medical history, and via linkages to Hospital Episode Statistics, national cancer registry, the Office for National Statistics death register, follow-up assessments where repeat measurements were taken, imaging, and genetic sequencing. It therefore offers a potential data source for an external validation of models developed in **Chapters 3 & 5,** and for the evaluation of the incremental effects on model performance of including additional data types. Examples of these include comparative validation of clinical prediction models for colorectal and kidney cancers[5,6], and assessing the effects of integrating genomic risk scores into models for coronary artery disease[7]. However, the UK Biobank has a narrower age range of recruited women compared to the development cohorts from QResearch, is affected by

healthy volunteer bias, and reduced ethnic diversity compared to the wider population[8,9], which should be considered.

Factors related to reproductive history are included in several extant breast cancer prediction models[10], but were not recorded well enough in the QResearch data to be included as predictors. Other models have incorporated genetic risk, such as integrating the output of a polygenic risk scores (PRSs) [11,12] or as standalone PRS[13-15]. These are summations of the genetic liability to a trait, typically using single nucleotide polymorphisms (SNPs) distributed throughout the genome[16]. Integrating genetic information may improve model discrimination[11,17], but the effects vary across diseases studied[7,18], and again, as discussed in **Chapter 5**, the distinction between incident cancer and life-threatening cancer may be relevant. In a Swedish case-only study of 5,500 individuals with breast cancer, women at higher risk based on the Tyrer-Cuzick model or a 77-SNP PRS were more likely to be diagnosed with tumours with favourable pathological factors, such as ER positivity, and were less likely to be diagnosed with metastatic tumours[19]. Conversely, risk estimated with a 313-SNP PRS was associated with increased risk of breast cancer diagnosis, and poor prognosis breast cancer (e.g. defined as metastatic, lymph node positive or triple-negative)[20]. Clinical prediction modelling studies tend to report the effects of PRS integration in terms of overall/average effects on performance metrics and ignore calibration[14], whereas the potential heterogeneity in benefits/harms such as in different sub-groups, is under-researched. This is despite prevalent concerns regarding the poor ethnic representation in the samples used to develop many PRSs[16].

Larger PRSs that integrate the effects of SNPs across the genome without selecting based on a 'significant' effect size threshold tend to have better performance. Recent work by

Genomics plc used Bayesian approaches to estimate non-zero weights for over 6 million SNPs to generate PRSs for 53 traits, including breast cancer diagnosis[21]. These were developed by meta-analysing results from genome-wide association studies, calculated for UK Biobank participants (after standard genetic data quality control[16,21]), and are available for use by UK Biobank-approved researchers.

An external evaluation of the Cox and competing risks regression models for incident breast cancer diagnosis and breast cancer mortality, respectively, are undertaken in the UK Biobank. Thereafter, the incremental effects of integrating the output from the Genomics plc breast cancer PRS, and reproductive factors are assessed. The models developed in **Chapter 5** refer to breast cancer mortality, therefore this represents an evaluation of the PRS for a related or 'proxy' trait. An alternative, specific breast cancer-mortality PRS was not available, and other studies have failed to find any germline variants associated with breast cancer mortality in a large, pooled cohort of women diagnosed with the condition[22]. Exploring the integration of polygenic risk to the prognostic models in **Chapter 4** using UK Biobank is not performed due to the systematic missingness of many predictors (such as receptor sub-types).

Data sources and study cohort derivation

All female participants enrolled in UK Biobank were identified (self-reported female sex). Females with a previous recorded diagnosis of breast cancer or breast carcinoma *in situ* prior to cohort entry were excluded (on self-report at the baseline assessment, or cancer registry data). Those with incomplete predictor variable data for the QResearch models were excluded, as were individuals that did not have a polygenic risk score

estimate from the 'standard' PRS of Thompson, et al. (assumed due to the individuals' obtained genetic data not pass standard quality control). This yielded a single study cohort suitable for external evaluation and integrated modelling.

Outcome definitions

The prediction date and start of follow-up was the date of UK Biobank assessment centre attendance. Follow-up was calculated from this date to the earliest of: date of event of interest occurring, or censoring (earliest of: reached 10-years follow-up alive, died of other cause prior to study end, or reached censoring date). The censoring date was 30th September 2021, informed by the dates up to which data were reported in death registers, the cancer registry HES, and GP data at the time of data extract.

Incident breast cancer diagnosis was defined as presence of a relevant ICD-10 code in the linked general practice, HES, cancer registry (ICD-10 or ICD-9), or ONS death certificate data linked to UK Biobank. Breast cancer death was defined as relevant ICD-10 codes recorded on death certification as either the primary or secondary (contributory) cause of death. Death from other, non-breast cancer-related causes was a competing risk which was accounted for when appropriate.

Crude event rates per 10,000 person-years (overall, and per 5-year age-groups) were estimated for breast cancer incidence and breast cancer mortality in the final analysis cohort. These were descriptively compared with corresponding data from the QResearch model derivation cohorts, and national statistics obtained from Cancer Research UK[23].

Predictor parameters

Predictions were generated using data available at the time of assessment centre attendance (prediction time), with individual predictor values ascertained by combinations of the UK Biobank assessment centre and the other linked data sources.

A code-mapping exercise was undertaken between the final model variables and relevant UK Biobank data fields. These and the endpoint variable definitions in terms of UK Biobank data fields are summarised in **Table 6.1**.

Parameter	UKB baseline assessment centre records	GP records	Death registry	HES	Cancer registry
Age (at start of follow-up)	Age when attended assessment centre , p21003_i0				
Biological sex	Sex p31				
Start of follow-up	Date of attending assessment centre, p53_i0				
Assessment centre (for IECV)	UK Biobank assessment centre p54_i0				
Body mass index at baseline	Body mass index (BMI), p21001_i0				
Townsend deprivation score	Townsend deprivation index at recruitment, p189				
Ethnic background	Ethnic background, p21000_i0				
Date lost to follow-up	Date lost to follow-up, p191				
Reason lost to follow-up	Reason lost to follow up, p190				
Date of death (1st instance)			Date of death p40000_i0		
Date of death (2nd instance – UKB has some			Date of death p40000_i1		

additional death records not captured in the above)					
Primary cause of death (Breast cancer = ICD10 code C50.9)			Primary cause of death on certificate p40001_i0		
Contributory causes of death (Breast cancer = ICD10 code C50.9)			Contributory cases of death on certificate P40002_i0 (arrays 1-9)		
Breast cancer, breast carcinoma in situ (e.g. DCIS), lung cancer, thyroid cancer, haematological cancer, gynaecological cancers diagnosed Cancer-specific ICD10 and ICD-9 codes are detailed in the data processing code	Previous self-reported cancer, p20001_i0	Data field 42040 (GP clinical event records)	Primary cause of death on certificate p40001_i0 Contributory cases of death on certificate P40002_i0 (arrays 1-9)	Relevant ICD10 codes in p41270 columns	Relevant ICD-10 codes in: p40006_i0, p40006_i1, p40006_i2, p40006_i3, p40006_i4, p40006_i5, p40006_i6 Relevant ICD9 codes in: p40013_i0, p40013_i0, p40013_i1, p40013_i2, p40013_i3, p40013_i4, p40013_i5, p40013_i6 (Accompanying dates: p40005_i0, p40005_i1, p40005_i2, p40005_i3, p40005_i4, p40005_i5, p40005_i6)

Polygenic risk scores ('relative risk')	p26220 (prs_standard) p26221 (prs_enhanced)	
Age at menarche	p2714_i0	
Number of live births	p2734_i0	
Age at first live birth	p2754_i0	
Ever used oral contraceptive pill	p2784_i0	
Hysterectomy	p3591_i0	
Smoking status	p20116_i0 (current smoking status), p3456_i0 (number of cigarettes smoked per day) Never smoker if p20116_i0 = "Never"; ex-smoker if p20116_i0 = "Previous"; light, moderate and heavy smoking based on current smoker = 1, and number of cigs per day (p3456_i0)	

Alcohol intake	p20117_i0 (current alcohol intake status); p1558_i0 (alcohol intake frequency); p1568_i0 (average weekly red wine intake); p1578_i0 (average white wine and champagne intake); p1588_i0 (average weekly beer and cider intake); p1598_i0 (average weekly spirits intake); p1608_i0 (average weekly fortified wine intake); p5364_i0 (average weekly other alcoholic drinks intake)		
Polycystic ovarian syndrome, benign breast disease	p20002_i0	Read codes in data field 42040 (GP clinical event records)	
Family history of		Read codes in data field 42040	

gynaecological cancer, vasculitis		(GP clinical event records)	
Hormone replacement therapy, selective serotonin reuptake inhibitor, anti-psychotic use		Data field 42039 (GP prescription records)	
Family history of breast cancer	Illnesses of mother p20110, illnesses of siblings p20111		
	First occurrences data fields (first recorded instances in any of GP records, UK Biobank baseline assessment centre, HES or death register), with relevant ICD10 codes in brackets		
Type 1 diabetes	p130706 (E10)		
Type 2 diabetes	p130708 (E11)		
Ischaemic heart disease			
Psychotic illness	p131296 (I20); p131298 (I21); p131300 (I22); p131304 (I24); p131306 (I25) p130874 (F20); p130876 (F21); p130878 (F22); p130880 (F23); p130884 (F25); p130886 (F28); p130890 (F30); p130892 (F31)		
Endometriosis	p132122 (N80)		

Table 6.1. Data dictionary demonstrating the definition of model predictor variables (QResearch 'phenotypic' models), additional parameters of interest for inclusion in integrated modelling, and the outcomes of interest. Grey space corresponds to the lack of use of that data source for the parameter of interest. GP = general practice, HES = Hospital Episode Statistics, IECV = internal-external cross-validation.

Missing data

Absence of recorded diagnoses on UK Biobank assessment centre questionnaires, cancer registry or linked healthcare datasets was assumed to represent absence of that condition. There was data missingness for some predictors in the Endpoint 1 & 2 models, and for some of the reproductive history data fields.

UK Biobank participants self-selected into a prospective bio-banking study, and underwent phenotyping via extensive electronic questionnaires, and nurse-led, face-to-face interviews. On review of the UK Biobank Data Showcase and from personal previous experience of this dataset, the typical value assigned to non-recorded data (e.g. for missing smoking status) is "prefer not to answer". The missing not at random (MNAR) assumption is therefore deemed the likeliest data missingness mechanism, for which multiple imputation with chained equations is inappropriate[24-26], and the optimal approach to handle this scenario is unclear[24]. In this context, complete case analyses were conducted. Further, the missingness of polygenic risk score estimates is likely driven by characteristics of the genetic data itself rendering it unsuitable during quality control – the proclivity therefore of an individual to have missing genetic risk scores is likeliest to represent being missing not at random.

The work in UK Biobank comprised the external evaluation of two models developed earlier in this thesis, followed by integrated modelling, the methods for which are discussed in turn.

External evaluation of models in UK Biobank cohort

Ten-year risk estimates from both models were calculated for all eligible individuals in the final extracted cohort dataset. Both models were assessed separately due to the difference in outcome definitions. Both models were assessed in terms of discrimination (Harrell's C-index at 10-years) and calibration (calibration slope and calibration-in-the-large)[27]. These metrics were estimated for each assessment centre, and pooled using random effects meta-analysis (Hartung-Knapp-Sidik-Jonkmann method[28]) to provide an overall meta-estimate of performance and also explore heterogeneity and stability in model performance.

Smoothed calibration plots were generated by plotting the risk predictions against pseudo-observations for the Kaplan-Meier failure probability estimate of incident breast cancer at 10-years (Cox model, 'Endpoint 1')[29] or pseudo-observations for the Aalen-Johansen estimate of the cumulative incidence function for breast cancer mortality at 10-years (competing risks regression model, 'Endpoint 3')[30] with running smoothers. Decision curve analysis was also performed for both models (unadjusted for, and accounting for competing risks of non-breast cancer-related death)[31].

For the incident breast cancer model, Royston & Sauerbrei's D and R^2_D were estimated[32]. Harrell's C was estimated using inverse probability of censoring weights for the competing risks model (breast cancer mortality, 'Endpoint 3').

Recalibration of either model[33] was not initially performed as it was *a priori* deemed likely that the models would likely be miscalibrated on external validation in this selected cohort. Recalibration would not affect discrimination and updating the primary models to

be better calibrated to the selected UK Biobank cohort may yield a model with reduced applicability to the wider female population. This is explored in **Chapter 7**.

Integrated model – development and evaluation

Integrated modelling considered different combinations of the linear predictor (LP) from the phenotypic models with two reproductive factors (RFs; age at menarche, number of live births), the PRS output (reported as a relative risk), and interactions (*). Using the linear predictor from the breast cancer incidence and breast cancer mortality models provided an easy-to-use summation of risk from the routinely collected 'phenotypic' predictor variables – this could be then handled non-linearly if needed and interacted on other terms easily in integrated modelling steps.

The following models were fitted within Cox proportional hazards (breast cancer incidence modelling, 'Endpoint 1') and pseudo-observations-based competing risks frameworks (breast cancer mortality modelling, 'Endpoint 3'):

1) LP + RFs
2) LP + PRS + LP*PRS
3) LP + RF + PRS + LP*PRS

These models were fitted to the final cohort without any predictor selection. The LP, RFs and PRs were kept as continuous variables. Fractional polynomials were used to handle non-linearities for each of these, with up to 2 non-linear terms permitted.

Due to the smaller size of this dataset compared to the cohorts from QResearch, the risk of optimism is to be considered alongside plans for assessing model generalisability.

Bootstrapping with 500 repeats was used to calculate estimates of Harrell's C, calibration slope and calibration-in-the-large, their confidence intervals, and optimism-corrected performance metrics[2]. A uniform shrinkage factor based on the average calibration slope across all bootstraps was calculated and applied to model coefficients to produce the final models.

To estimate model generalisability and performance heterogeneity, internal-external cross-validation was then used using non-random geographical splitting based on the location of UK Biobank assessment centre, with random effects meta-analysis used to calculate an overall meta-estimate of model performance with 95% confidence intervals and 95% prediction intervals. As per the modelling in **Chapters 3-5**, the produced held-out predictions were used to estimate Harrell's C and summary calibration metrics overall and in ethnic sub-groups and generate smoothed calibration plots. Decision curves were plotted for each model to assess the incremental clinical utility overall of adding additional predictors for each endpoint (incidence and mortality).

Sample size calculations

External validation

For the external validation, the methods of Riley, et al.[34] were used to assess if the expected sample size from UK Biobank was sufficient. For the Cox model predicting 10-

year incident breast cancer, the standard deviation of the linear predictor distribution was calculated in women aged 40-70 years (age range of UK Biobank) in the QResearch dataset, as was the distribution of survival times. Censoring distributions sought to approximate those in UK Biobank. Assuming a validation set of 240,000 individuals, it was estimated across 500 iterations that the mean number of events would be 8,350, and the mean standard error for the calibration slope would be 0.03 (target < 0.1).

There is no clear best practice to estimating minimum required sample size for external validation of a competing risks model, and the aforementioned study highlighted potential complications of a competing risks approach [34]. An indicative sample size calculation was performed using the distribution of the linear predictor from the Cox model predicting 10-year breast cancer mortality and the corresponding distribution of survival times. Assuming a validation set of 240,000 individuals, it was estimated across 500 iterations that the mean number of events would be 638, and the mean standard error for the calibration slope would be 0.08 (target <0.1).

Development of integrated models

Regarding the development of the integrated models for breast cancer incidence: assuming 14 predictors (if 2 FP terms selected for each continuous variable, and two two-way interaction terms between FP terms for LP and PRS), an event rate of 0.0029 (2.09 per 1,000 person-years), a conservative 15% of the maximum permitted Cox-Snell R^2 (0.555) of 0.08325, a mean follow-up of 9.8 years derived from a recent study using UK

Biobank data, and a prediction horizon of 10 years, 1,443 individuals with 42 events was required (events per predictor [EPP] = 2.93)[35].

For the development of integrated models for breast cancer mortality, similar assumptions were used, but with 10% of the event rate (0.00029), a Cox-Snell R^2 of 0.000585 (15% of maximum, 0.0039), 215,315 individuals were required with 612 events (EPP = 43.71).

Assuming that single terms would be selected for each predictor variable, i.e. 6 predictor parameters, the required sample size for the incidence model would be 619 individuals with 18 events (EPP = 2.93), and for the mortality model it would be 92,278 individuals with 263 events (EPP = 43.71).

On review of the UK Biobank Data Showcase on 1st August 2022 (prior to data extraction), there were 20,223 cases of breast cancer recorded (data field 40006) albeit without being able to remove prevalent cases at baseline. There were 1,473 cases where breast cancer was primary cause of death (data field 40001) and 181 cases where it was a contributory cause of death (data field 40002).

Approval, conduct and analysis code

The UK Biobank Access Management Committee approved this project (reference 87818). Ethical approval for UK Biobank is from the North-West Multi-Centre Research Ethics Committee, reference: 16/NW/0274. Analyses were conducted using the UK Biobank's Research Access Platform using R and Stata V17, code is available at the following GitHub repository:

https://github.com/AshDF91/UKB_integrated_polygenic_modelling

Chapter references

1. Riley RD, Ensor J, Snell KI, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 2016; **353**: i3140.

2. Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol* 2016; **69**: 245-7.

3. Collins R. What makes UK Biobank special? *Lancet* 2012; **379**(9822): 1173-4.

4. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015; **12**(3): e1001779.

5. Harrison H, Pennells L, Wood A, et al. Validation and public health modelling of risk prediction models for kidney cancer using the UK Biobank. *BJU Int* 2022; **129**(4): 498-511.

6. Usher-Smith JA, Harshfield A, Saunders CL, et al. External validation of risk prediction models for incident colorectal cancer using UK Biobank. *Br J Cancer* 2018; **118**(5): 750-9.

7. Elliott J, Bodinier B, Bond TA, et al. Predictive Accuracy of a Polygenic Risk Score-Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease. *JAMA* 2020; **323**(7): 636-45.

8. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* 2017; **186**(9): 1026-34.

9. Tyrrell J, Zheng J, Beaumont R, et al. Genetic predictors of participation in optional components of UK Biobank. *Nat Commun* 2021; **12**(1): 886.

10. Zheng Y, Li J, Wu Z, et al. Risk prediction models for breast cancer: a systematic review. *BMJ Open* 2022; **12**(7): e055398.

11. Brentnall AR, Cuzick J, Buist DSM, et al. Long-term Accuracy of Breast Cancer Risk Assessment Combining Classic Risk Factors and Breast Density. *JAMA Oncol* 2018; **4**(9): e180174.

12. 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**(4): 476-82.

13. 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**(1): 21-34.

14. Mars N, Koskela JT, Ripatti P, et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat Med* 2020; **26**(4): 549-57.

15. Mars N, Widen E, Kerminen S, et al. The role of polygenic risk and susceptibility genes in breast cancer over the course of life. *Nat Commun* 2020; **11**(1): 6383.

16. Choi SW, Mak TS, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc* 2020; **15**(9): 2759-72.

17. Evans DGR, van Veen EM, Harkness EF, et al. Breast cancer risk stratification in women of screening age: Incremental effects of adding mammographic density, polygenic risk, and a gene panel. *Genet Med* 2022; **24**(7): 1485-94.

18. Olsen M, Fischer K, Bossuyt PM, et al. Evaluating the prognostic performance of a polygenic risk score for breast cancer risk stratification. *BMC Cancer* 2021; **21**(1): 1351.

19. Holm J, Li J, Darabi H, et al. Associations of Breast Cancer Risk Prediction Tools With Tumor Characteristics and Metastasis. *J Clin Oncol* 2016; **34**(3): 251-8.

20. McCarthy AM, Manning AK, Hsu S, et al. Breast cancer polygenic risk scores are associated with short-term risk of poor prognosis breast cancer. *Breast Cancer Res Treat* 2022; **196**(2): 389-98.

21. Thompson DJ, Wells D, Selzam S, et al. UK Biobank release and systematic evaluation of optimised polygenic risl scores for 53 diseases and quantitative traits. *medRxiv* 2022: https://www.medrxiv.org/content/10.1101/2022.06.16.22276246v2.

22. Escala-Garcia M, Guo Q, Dork T, et al. Genome-wide association study of germline variants and breast cancer-specific mortality. *Br J Cancer* 2019; **120**(6): 647-57.

23. Cancer Research UK. Breast cancer statistics. 2022. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer.

24. Hughes RA, Heron J, Sterne JAC, et al. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *Int J Epidemiol* 2019; **48**(4): 1294-304.

25. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009; **338**: b2393.

26. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med* 2011; **30**(4): 377-99.

27. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010; **21**(1): 128-38.

28. 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.

29. Austin PC, Harrell FE, Jr., van Klaveren D. Graphical calibration curves and the integrated calibration index (ICI) for survival models. *Stat Med* 2020; **39**(21): 2714-42.

30. Austin PC, Putter H, Giardiello D, et al. Graphical calibration curves and the integrated calibration index (ICI) for competing risk models. *Diagn Progn Res* 2022; **6**(1): 2.

31. Van Calster B, Wynants L, Verbeek JFM, et al. Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. *Eur Urol* 2018; **74**(6): 796-804.

32. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004; **23**(5): 723-48.

33. Booth S, Riley RD, Ensor J, Lambert PC, et al. Temporal recalibration for improving prognostic model development and risk predictions in settings where survival is improving over time. *Int J Epidemiol* 2020; **49**(4): 1316-25.

34. Riley RD, Collins GS, Ensor J, et al. Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome. *Stat Med* 2022; **41**(7): 1280-95.

35. Riley RD, Snell KI, 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**(7): 1276-96.

Chapter Seven

Results from external evaluation and integrated clinical and phenotypic modelling in UK Biobank

Summary

This chapter presents and discusses the results from two sets of analyses using data from UK Biobank – an external evaluation of two new models developed using QResearch and linked datasets (for incident breast cancer diagnosis [**Chapter 3**], and breast cancer mortality in women without breast cancer at baseline [**Chapter 5**]), and an assessment of the effects of integrating polygenic and reproductive history information to each of these models.

Study populations

Following data processing and the exclusions discussed in **Chapter 6**, a total of 242,996 female individuals were eligible for inclusion in these cohort analyses from a potential pool of 273,324 females (**Figure 7.1**).

From the perspective of the incident breast cancer diagnosis model, within a total of 2,956,114 person-years of follow-up (maximum 15.55 years) there were 10,030 incident

breast cancer diagnoses recorded, yielding a crude incidence rate of 33.93 per 10,000 person-years (95% CI: 33.27-34.60). The ascertainment yield of using different linkages within UK Biobank is summarised in **Table 7.1**. The overall crude incidence rate was higher than that observed in the derivation dataset (20.59 per 10,000, 95% CI: 20.48-20.70). Due to different cohort age ranges, age group-stratified crude rates were compared between the two datasets and publicly available population data from Cancer Research UK (CRUK, **Table 7.2**). Data from QResearch appeared broadly similar to population breast cancer incidence rates across most strata, and QResearch and UK Biobank appeared comparable for women of screening age. All data sources showed decreased incidence rates in women aged 70-74 years compared to the immediately preceding groups, which may partly reflect the cessation of routine screening mammography offered in the UK's NHS Breast Screening Programme at 70 years of age. When restricting to 10-years of follow-up (the model prediction horizon) there were 8,243 events during 2,357,613 person-years in the UK Biobank cohort (crude incidence rate 34.95, 95% CI: 34.22 to 35.72).

There were 663 breast cancer-related deaths within 3,017,874 person-years of follow-up (maximum 15.55 years), yielding a crude mortality rate of 2.20 per 10,000 person-years (95% CI: 2.04 to 2.37). When comparing age group-stratified mortality rates between the final UK Biobank study cohort and QResearch, rates appeared lower for the former with non-overlapping confidence intervals (**Table 7.3**). The QResearch event rates were slightly lower than the CRUK national figures, which may be due to different exclusion criteria, i.e. the model derivation study excluded prevalent breast cancers, whereas this was not performed for the national statistics. When restricting to 10-years follow-up, there were 420 breast cancer-related deaths (5.4% of all deaths) within 2,396,841 person-years

(crude mortality rate 1.75 per 10,000 person-years, 95% CI: 1.59 to 1.93). There were 7,333 non-breast cancer-related deaths recorded (3.02% of cohort).

External validation and integrated modelling for both models used geographical splitting by UK Biobank assessment centre, therefore crude incidence and mortality rates by centre were calculated (**Table 7.4 & 7.5**). Due to low event counts in some individual assessment centres that may affect precision or require suppression, and to have consistent geographical 'units' in the evaluation framework, some centres were grouped based on geographical proximity. Data for Cardiff, Swansea and Wrexham were grouped to form a 'Wales' unit, and Manchester and Stockport centres were combined.

The baseline characteristics in terms of the predictors included in both models are summarised in **Table 7.6 –** these are presented for the final UK Biobank study cohort and the derivation cohort from QResearch. Whilst not a formal statistical comparison, there were some interesting differences in crude percentages/distributions of baseline predictors between the derivation data from QResearch and those obtained from UK Biobank. For example, women in UK Biobank reported an 81.37% rate of ever using oral contraceptive pill compared to the 11.81% of women as recorded in the QResearch model derivation data. This likely reflects the differences in methods of data collection, i.e. recall prompted by a questionnaire in an opt-in biobanking scenario may yield higher completeness compared to relying on coded/recorded data in longitudinal electronic healthcare records, particularly when women may have ceased using oral contraception many years ago. Another example is the difference in recorded family history of breast cancer (10.68% in UK Biobank versus 1.53% in QResearch data). This again may reflect recall and data collection approaches, rather than 'true' differences in the risk of the different study samples.

This chapter did not seek to wholly refit the models and re-estimate all model coefficients. Due to the size and population-representativeness of QResearch, using this data source would be likelier to yield a model that is generalisable to the target (wider) female population. The UK Biobank has different selection mechanisms and different predictor distributions. The latter could manifest as different predictor-risk associations if a model were to be refitted.

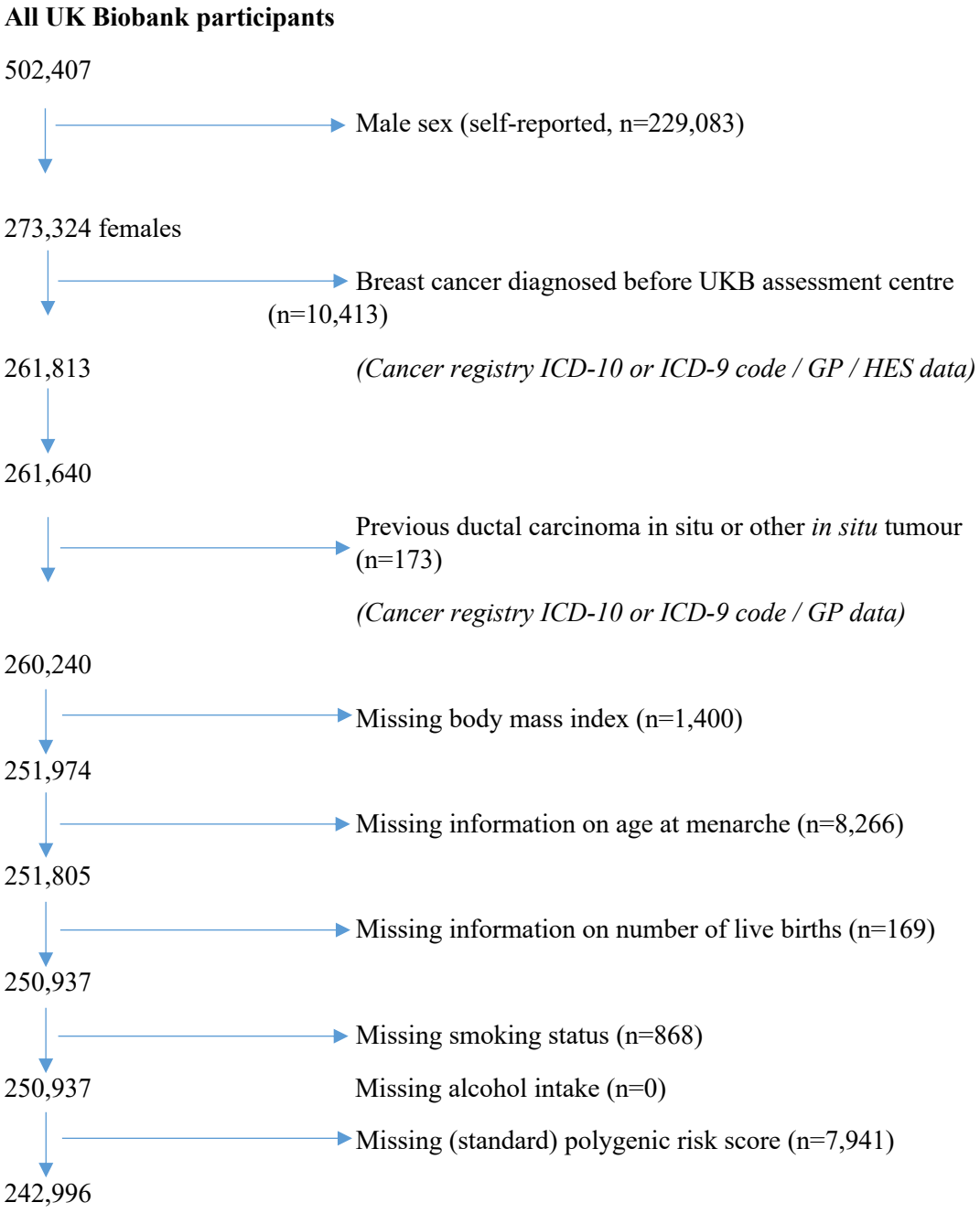

Figure 7.1. Flow chart summarising the exclusions to derive the final cohort dataset for external validation and integrated modelling in UK Biobank. ICD = International Classification of Diseases, GP = general practice data, HES = Hospital Episodes Statistics.

Data source(s)	Number of incident breast cancers	Person-years	Crude incidence rate per 10,000 (95% CI)
UKB: Cancer registry only	8,680	2,960,613	29.32 (28.71 to 29.94)
UKB: Cancer registry + HES	9,843	2,957,689	33.28 (32.63 to 33.94)
UKB: Cancer registry + HES + GP	10,017	2,956,114	33.89 (33.23 to 34.56)
UKB: Cancer registry + HES + GP + deaths register	10,030	2,956,114	33.93 (33.27 to 34.60)
Data source	**Number of incident breast cancers**	**Person-years**	**Crude incidence rate per 10,000 (95% CI)**
QResearch (model derivation study; GP + HES + Cancer registry + deaths register)	142,712	69,310,127	20.59 (20.48 to 20.70)

Table 7.1. Crude breast cancer incidence rates in UK Biobank (UKB, and different combinations of linked databases) compared to that observed in the QResearch (plus linked databases) cohort used to initially derive the phenotypic models, as detailed in **Chapter 3**. HES – hospital episodes statistics, GP = general practice records, CI = confidence interval.

	Breast cancer incidence rate per 10,000 (95% CI)		
Age group	**UK Biobank study cohort**	**QResearch study cohort**	**Cancer Research UK statistics**
40-44 years	15.33 (12.67 to 18.54)	11.56 (11.31 to 11.82)	12.5
45-59 years	30.64 (28.43 to 33.03)	20.07 (19.73 to 20.42)	21.5
50-54 years	29.41 (27.74 to 31.17)	29.99 (29.56 to 30.43)	28.0
55-59 years	29.47 (27.96 to 31.06)	29.59 (29.14 to 30.05)	28.5
60-64 years	36.32 (34.79 to 37.92)	36.41 (35.88 to 36.95)	33.8
65-69 years	41.34 (39.75 to 42.99)	39.71 (39.11 to 40.31)	41.2
70-74 years	33.46 (31.79 to 35.21)	35.92 (35.32 to 36.54)	37.3
75-79 years	34.24 (31.59 to 37.12)	37.68 (37.00 to 38.37)	40.3
80-85 years	43.59 (34.81 to 54.58)	40.43 (39.64 to 41.22)	43.0

Table 7.2. Age group-specific breast cancer incidence rates for women in the study cohort from UK Biobank (and linked datasets) compared to those observed in the QResearch (and linked datasets) cohort used to initially derive the phenotypic models, and also national data reported by Cancer Research UK covering years 2017-2019. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/incidence-invasive#heading-One (accessed 23rd Sept 2022).

	Breast cancer mortality rate per 10,000 (95% CI)		
Age group	**UK Biobank study cohort**	**QResearch study cohort**	**Cancer Research UK statistics**
40-44 years	Not estimable	1.07 (0.99 to 1.15)	1.3
45-59 years	1.24 (0.86 to 1.80)	1.75 (1.65 to 1.85)	2.3
50-54 years	1.23 (0.93 to 1.63)	2.48 (2.36 to 2.61)	3.3
55-59 years	1.71 (1.37 to 2.12)	3.13 (2.99 to 3.28)	4.0
60-64 years	1.60 (1.31 to 1.96)	3.77 (3.60 to 3.95)	4.8
65-69 years	2.16 (1.82 to 2.56)	4.78 (4.58 to 4.99)	5.8
70-74 years	3.26 (2.77 to 3.82)	6.53 (6.28 to 6.79)	8.1
75-79 years	6.25 (5.20 to 7.53)	9.23 (8.99 to 9.57)	10.7
80-85 years	9.88 (6.23 to 15.68)	12.72 (12.30 to 13.17)	15.4

Table 7.3. Age group-specific breast cancer mortality rates for women in the study cohort from UK Biobank (and linked datasets) compared to those observed in the QResearch (and linked datasets) cohort used to initially derive the phenotypic models, and also national data reported by Cancer Research UK covering years 2017-2019. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/mortality#heading-One (accessed 23rd Sept 2022). An exclusion criterion for the UK Biobank and QResearch final study cohorts was presence of previously diagnosed breast cancer – the national statistics do not make this exclusion. Therefore, the trajectory is slightly different to the national statistics, however, it is apparent that women in the UK Biobank had lower age-group mortality rates.

Assessment centre	Number of events	Total follow-up (person-years)	Crude rate per 10,000 (95% CI)
Barts (London)	217	60,950.36	35.60 (31.17 to 40.67)
Birmingham	401	112,918.77	35.51 (32.20 to 39.16)
Bristol	776	210,408.49	36.88 (34.37 to 39.57)
Bury	469	131,243.66	35.74 (32.64 to 39.12)
Wales	261	98,922.48	26.38 (23.37 to 29.79)
Croydon	447	130,422.86	34.27 (31.24 to 37.62)
Edinburgh	270	84,072.14	32.12 (28.50 to 36.18)
Glasgow	311	90,373.94	34.41 (30.79 to 38.46)
Hounslow	460	132,655.82	34.68 (31.64 to 37.99)
Leeds	737	210,860.78	34.95 (32.52 to 37.57)
Liverpool	588	154,161.51	38.14 (35.18 to 41.35)
Manchester & Stockport	198	64,884.21	30.52 (26.55 to 35.08)
Middlesbrough	350	99,741.52	35.09 (31.60 to 38.97)
Newcastle	632	177,198.52	35.67 (32.99 to 38.56)
Nottingham	556	161,044.55	34.52 (31.77 to 37.52)
Oxford	237	71,277.23	33.25 (29.28 to 37.76)
Reading	554	140,618.31	39.40 (36.25 to 42.82)
Sheffield	499	141,528.23	35.26 (32.30 to 38.49)
Stoke	280	84,329.76	33.20 (29.53 to 37.33)

Table 7.4. UK Biobank assessment centre-specific breast cancer incidence rates per 10,000 person-years. The 'Wales' group comprises Cardiff, Swansea and Wrexham assessment centres.

Assessment centre	Number of events	Total follow-up (person-years)	Crude rate per 10,000 (95% CI)
Barts (London)	9	69,156.71	1.45 (0.76 to 2.79)
Birmingham	21	114,735.20	1.83 (1.19 to 2.81)
Bristol	35	214,104	1.63 (1.17 to 2.28)
Bury	28	133,397.06	2.10 (1.45 to 3.04)
Wales	19	100,304.85	1.89 (1.21 to 2.97)
Croydon	27	132,498.65	2.04 (1.40 to 2.97)
Edinburgh	15	85,372.41	1.76 (1.06 to 2.91)
Glasgow	15	91,890.06	1.63 (0.98 to 2.71)
Hounslow	20	134,818.39	1.48 (0.96 to 2.30)
Leeds	37	214,402.81	1.73 (1.25 to 2.38)
Liverpool	35	157,010.10	2.23 (1.60 to 3.10)
Manchester & Stockport	10	65,838.74	1.52 (0.82 to 2.82)
Middlesbrough	16	101,444.33	1.58 (0.97 to 2.57)
Newcastle	22	180,221.15	1.22 (0.80 to 1.85)
Nottingham	24	163,738.47	1.47 (0.98 to 2.19)
Oxford	16	72,384.69	2.21 (1.35 to 3.61)
Reading	35	143,238.14	2.44 (1.75 to 3.40)
Sheffield	23	143,784.03	1.60 (1.06 to 2.41)
Stoke	13	85,701.64	1.52 (0.88 to 2.61)

Table 7.5. UK Biobank assessment centre-specific breast cancer mortality rates per 10,000 person-years. The 'Wales' group comprises Cardiff, Swansea and Wrexham assessment centres.

Parameter	Category	UK Biobank (Column %)	QResearch (Column %)
Total women		242,996	11,626,969
Age at entry (years)	Mean (SD)	56.20 (8.01)	41.78 (18.13)
	Median (IQR)	57 (50-63)	36 (27–53)
Body mass index at entry (kg/m^2)	Mean (SD)	27.07 (5.19)	25.37 (5.46)
	Median (IQR)	26.09 (23.43 to 29.70)	24.2 (21.5–28.1)
Townsend deprivation score	Mean (SD)	-1.36 (3.02)	0.71 (3.23)
Ethnic group	White	229,984 (94.65%)	6,168,419 (53.05%)
	South Asian	3,140 (1.29%)	507,829 (4.37%)
	Other Asian	725 (0.30%)	189,635 (1.63%)
	Black	3,862 (1.59%)	445,720 (3.83%)
	Chinese	862 (0.35%)	132,583 (1.14%)
	Other (inc. mixed race, Arab)	3,825 (1.57%)	328,396 (2.82%)
	Not recorded*	600 (0.25%)	3,854,387 (33.15%)
Smoking status	Non-smoker	144,718 (69.56%)	5,045,101 (43.39%)
	Ex-smoker	76,501 (31.48%)	1,445,584 (12.43%)
	Light smoker (1-9/day)	3,979 (1.64%)	1,318,132 (11.34%)
	Moderate smoker (10-19/day)	7,528 (3.10%)	308,372 (2.65%)
	Heavy smoker (20+/day)	10,270 (4.23%)	133,108 (1.14%)
	Not recorded	0	3,376,672 (29.04%)
Alcohol intake	Non-drinker	22,551 (9.28%)	4,120,142 (35.44%)
	Trivial <1u/day	112,764 (46.41%)	1,580,548 (13.59%)
	Light 1-2u/day	58,738 (24.17%)	601,071 (5.17%)
	Moderate 3-6u/day	45,876 (18.88%)	246,366 (2.12%)
	Heavy 7-9u/day	2,465 (1.01%)	9,823 (0.08%)
	Very Heavy >9u/day	602 (0.25%)	17,467 (0.15%)
	Not recorded	0	5,051,552 (43.45%)
Benign breast disease		2,701 (1.11%)	282,663 (2.43%)

Endometriosis		7,593 (3.12%)	151,158 (1.30%)
Polycystic ovarian syndrome		153 (0.06%)	197,886 (1.70%)
Hysterectomy		17,258 (7.10%)	99,439 (0.86%)
Previous gynaecological cancer		2,052 (0.84%)	26,626 (0.23%)
Oral contraceptive pill use (ever)		197,728 (81.37%)	1,372,633 (11.81%)
Recent oestrogen-only hormone replacement therapy	None (reference)	238,964 (98.34%)	11,467,510 (98.63%)
(<5 years since last prescription)	<1 year duration	604 (0.25%)	58,156 (0.50%)
	1 - 2.9 years duration	565 (0.23%)	34,566 (0.30%)
	3 – 4.9 years duration	550 (0.23%)	25,760 (0.22%)
	5 – 9.9 years duration	1,349 (0.56%)	30,254 (0.26%)
	10+ years duration	964 (0.40%)	10,723 (0.09%)
Past oestrogen-only hormone replacement therapy	None (reference)	239,498 (98.56%)	11,551,573 (99.35%)
(5+ years since last prescription)	<1 year duration	1,019 (0.42%)	35,352 (0.30%)
	1 - 2.9 years duration	672 (0.28%)	12,685 (0.11%)
	3 – 4.9 years duration	596 (0.25%)	8,732 (0.08%)
	5 – 9.9 years duration	991 (0.41%)	13,071 (0.11%)
	10+ years duration	220 (0.09%)	5,556 (0.05%)
Recent combined hormone replacement therapy	None (reference)	235,367 (96.86%)	11,339,053 (97.52%)
(<5 years since last prescription)	<1 year duration	1,140 (0.47%)	85,664 (0.74%)
	1 - 2.9 years duration	1,017 (0.42%)	68,532 (0.59%)
	3 – 4.9 years duration	1,064 (0.44%)	52,565 (0.45%)
	5 – 9.9 years duration	2,545 (1.05%)	63,127 (0.54%)
	10+ years duration	1,863 (0.77%)	18,028 (0.16%)

Past combined hormone replacement therapy	None (reference)	233,056 (95.91%)	11,489,012 (98.81%)
(5+ years since last prescription)	<1 year duration	3,063 (1.26%)	48,225 (0.41%)
	1 - 2.9 years duration	1,995 (0.82%)	26,529 (0.23%)
	3 – 4.9 years duration	1,617 (0.67%)	20,172 (0.17%)
	5 – 9.9 years duration	2,689 (1.11%)	31,508 (0.27%)
	10+ years duration	576 (0.24%)	11,523 (0.10%)
Family history of breast cancer		25,942 (10.68%)	177,368 (1.53%)
Family history of gynaecological cancer		51 (0.02%)	36,932 (0.32%)
Previous lung cancer		168 (0.07%)	9,414 (0.08%)
Previous haematological cancer		997 (0.41%)	31,637 (0.27%)
Previous thyroid cancer		397 (0.16%)	6,009 (0.05%)
Type 1 diabetes mellitus		958 (0.39%)	20,479 (0.18%)
Type 2 diabetes mellitus		4,268 (1.76%)	311,725 (2.68%)
Ischaemic heart disease		7,449 (3.07%)	255,299 (2.20%)
Vasculitis		229 (0.09%)	63,329 (0.54%)
Psychotic condition		1,233 (0.51%)	87,334 (0.75%)
Anti-psychotic medication use (ever)		3,124 (1.29%)	119,285 (1.03%)
Age at menarche (years)	Mean (SD)	12.97 (1.62)	[Not available]
	Median (IQR)	13 (12-14)	
Parity	Mean (SD)	1.83 (1.20)	[Not available]
	Median (IQR)	2 (1-2)	

Table 7.6. Summary characteristics of the final cohort for external validation and integrated risk prediction modelling obtained from UK Biobank, compared to the QResearch cohort in which the phenotypic models were developed. Due to the exclusion criteria for deriving the final UK Biobank cohort, there are no missing data for smoking or alcohol status. Missing data for UK Biobank individuals in terms of ethnic group reflects ethnicity not being a predictor in either model, but a variable of interest for model performance heterogeneity.

External evaluation

Model for 10-year risk of incident breast cancer diagnosis (Cox proportional hazards)

Predicted 10-year risks for the UK Biobank cohort ranged between 0.003 and 0.272 (mean predicted 10-year probability 0.023, median 0.021, IQR 0.015 to 0.028, see **Figure 7.2**).

The Cox proportional hazards model had a meta-analysis pooled Harrell's C statistic of 0.551 (95% CI: 0.543 to 0.559, **Figure 7.3**). The pooled calibration slope was 0.917 (95% CI: 0.902 to 0.932), and pooled calibration-in-the-large was -0.083 (95% CI: -0.098 to -0.068 which suggests slight overprediction). The smoothed calibration plot was notable for some under-prediction in low-risk individuals and over-prediction in those at highest risk (**Figure 7.4**).

Decision curve analysis demonstrated that the net benefit of this model was worse than 'screen all' strategies (**Figure 7.5**). To assess the extent to which this poor performance in the UK Biobank is driven by miscalibration, the baseline survival function was updated by fitting a *de novo* Cox model in the UK Biobank cohort with the model's linear predictor as a single variable and re-estimating the baseline survival at 10 years[1]. With this, net benefit improved to be similar to the 'screen all' strategy (**Figure 7.5**).

Applying a model to a dataset less heterogeneous than the derivation cohort, particularly one with a narrower age range in the context of age being the strongest predictor, can be expected to demonstrate lower discrimination – for context, the IECV-estimated Harrell's C statistic for screening-age (50-70 years) women in model development was 0.556 (95% CI: 0.550 to 0.562). Whilst the UK Biobank cohort recruited individuals aged 40 and above, the age distribution at baseline is concentrated between 50 and 70. These

discrimination indices are not dissimilar to results within other (phenotypic) models in similar age ranges[2,3], which did not report decision curves. Therefore, even after model recalibration to this 'target' setting, the integration of additional predictors beyond routinely collected healthcare variables may be needed to attain high-enough model discrimination to support risk-based screening. This is particularly pertinent if these models would be applied only to those within the age range currently eligible for the NHS Breast Screening Programme – phenotypic data alone may not be adequate to stratify women in middle age in terms of incident breast cancer diagnosis risks.

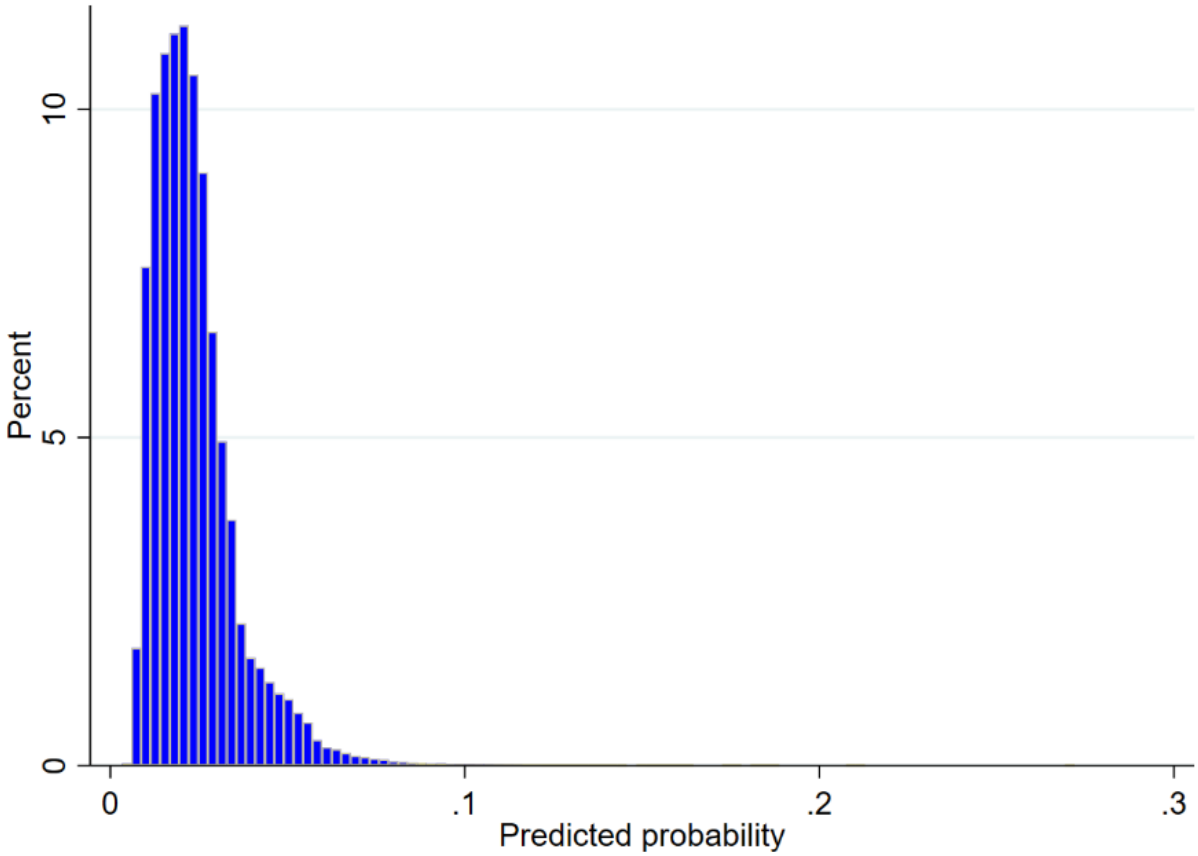

Figure 7.2. Distribution of predictions from Cox proportional hazards model for 10-year incident breast cancer diagnosis risk, estimated in the UK Biobank study cohort.

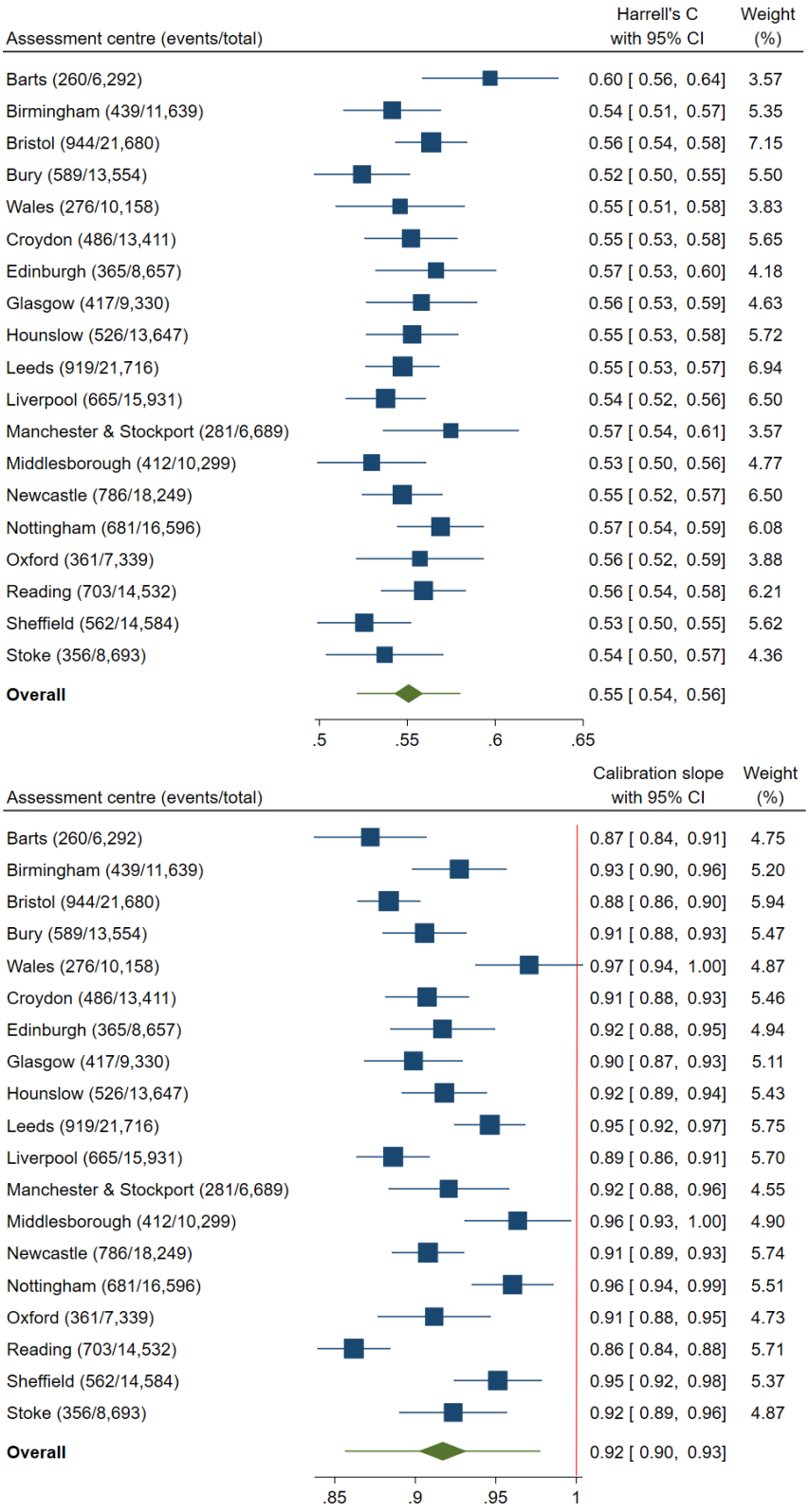

Figure 7.3. Meta-analysis forest plots demonstrating the assessment centre-level estimates, and random effects pooled meta-estimate for Harrell's C (top) and calibration slope (bottom) for the external evaluation of the Cox model predicting 10-year incident breast cancer diagnosis risk.

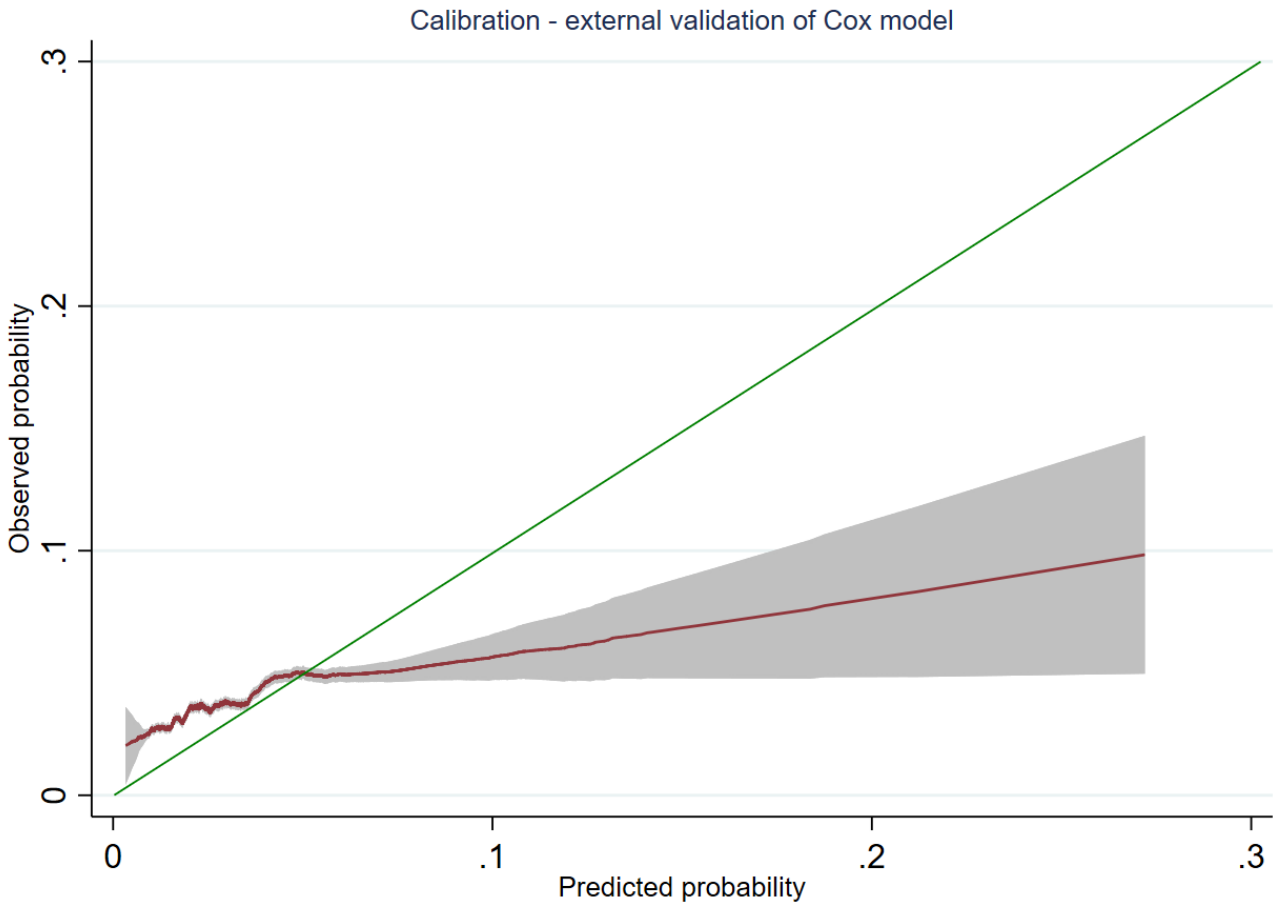

Figure 7.4. Smoothed calibration plot comparing predicted risks from the 10-year incident breast cancer diagnosis risk model, and the observed risks in the UK Biobank external evaluation cohort. Observed probabilities were obtained using pseudo-observations of the Kaplan-Meier failure function at 10 years.

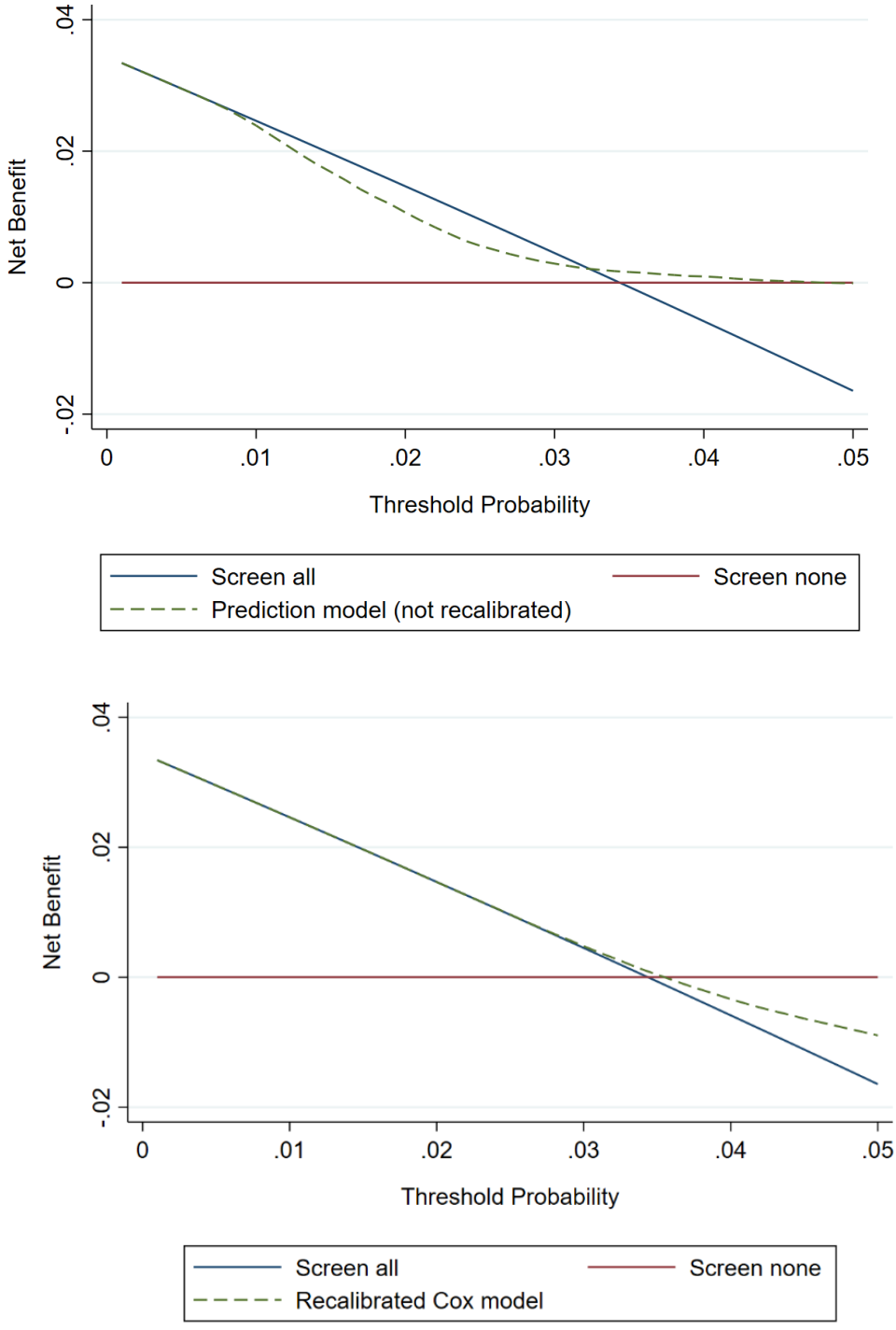

Figure 7.5. Decision curves for the Cox proportional hazards model predicting 10-year incident breast cancer diagnosis risk, applied to the UK Biobank cohort. The top demonstrates the results when applying the model without any recalibration; the bottom displays the results obtained with a Cox model recalibrated to the UK Biobank cohort by updating the model's baseline survival function.

Model for 10-year risk of breast cancer-related death (competing risks regression)

Predicted risks of 10-year breast cancer-related death in the UK Biobank cohort ranged between probabilities of 0.005 to 0.169 (mean predicted 10-year probability 0.004, median 0.003, IQR 0.003 to 0.005, **Figure 7.6**).

Due to the small numbers of deaths occurring in many assessment centre units, some regional performance metrics were estimated imprecisely. The meta-analysis pooled Harrell's C statistic was slightly higher than that observed for the incidence model, at 0.582 (95% CI: 0.546 to 0.621, **Figure 7.7**). The IECV-derived estimate for Harrell's C in the 40-70 years age group in the QResearch derivation data was 0.631 (95% CI: 0.613 to 0.650), and in the 50-70 years age group this was 0.579 (95% CI: 0.553 to 0.605, **Chapter 5**). The meta-analysis pooled calibration slope was 1.171 (95% CI: 1.151 to 1.191), and pooled calibration-in-the-large was 0.171 (95% CI: 0.151 to 0.191). The smoothed calibration plot demonstrated systematic over-estimation of risk (**Figure 7.8**), which aligns with the divergent crude mortality rates between the QResearch and UK Biobank cohorts. As QResearch had higher rates of the event of interest, the model fitted to that dataset might be expected to over-predict.

Decision curve analysis using the probabilities predicted by the competing risks regression model showed (very) minor improvement on net benefit compared to the 'screen all' strategy (**Figure 7.9**). Whilst recalibration was possible for the Cox model as aforementioned, the pseudo-observations based competing risk model does not have a baseline survival/cumulative incidence function and so cannot be recalibrated in the same way. Instead, a simple linear recalibration was performed by regressing the predicted probabilities from the competing risks model on pseudo-observations for the CIF for

breast cancer-related death at 10 years in the UK Biobank cohort – the predictions from this model were then used in decision curve analysis. After recalibration to the UK Biobank cohort, the model was associated with some improved net benefit compared to screen all strategy (**Figure 7.9**).

Whilst the performance of the competing risks model is overall better than the Cox model (both pre- and post-recalibration), there may still be benefit in integrating additional data types to further improve the model's utility in informing risk-based breast cancer strategies.

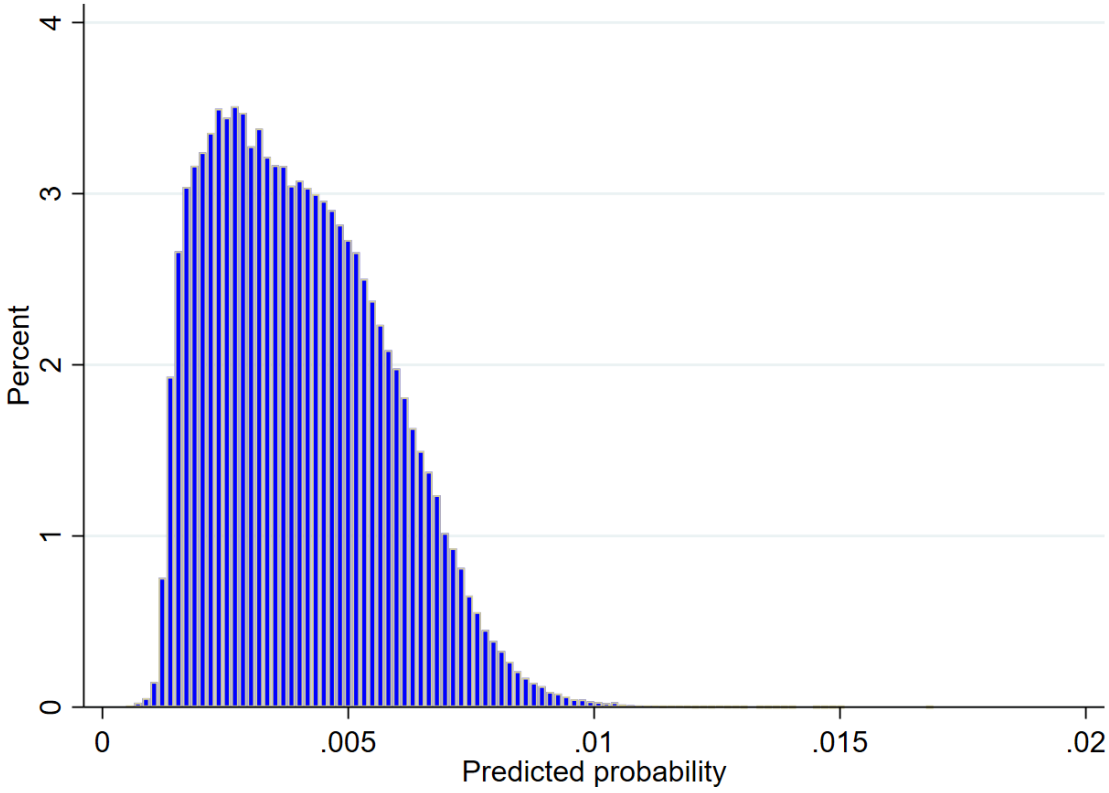

Figure 7.6. Distribution of predictions from the competing risks regression model for 10-year breast cancer-related death risk on application to the UK Biobank study cohort.

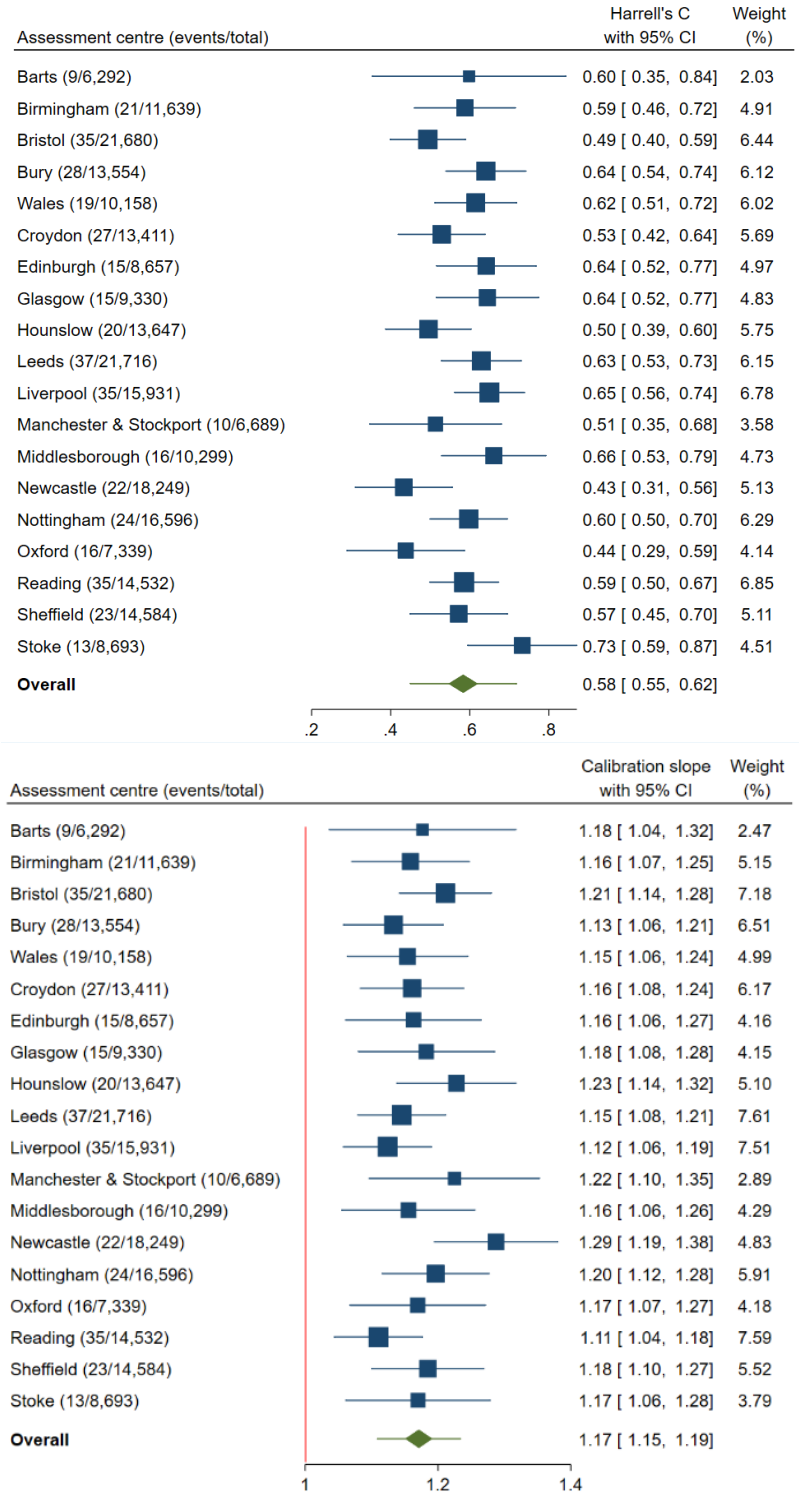

Figure 7.7. Meta-analysis forest plots demonstrating the assessment centre-level estimates, and random effects pooled meta-estimate for Harrell's C (top) and calibration slope (bottom) for the external evaluation of the competing risks regression model predicting 10-year breast cancer-related death risk.

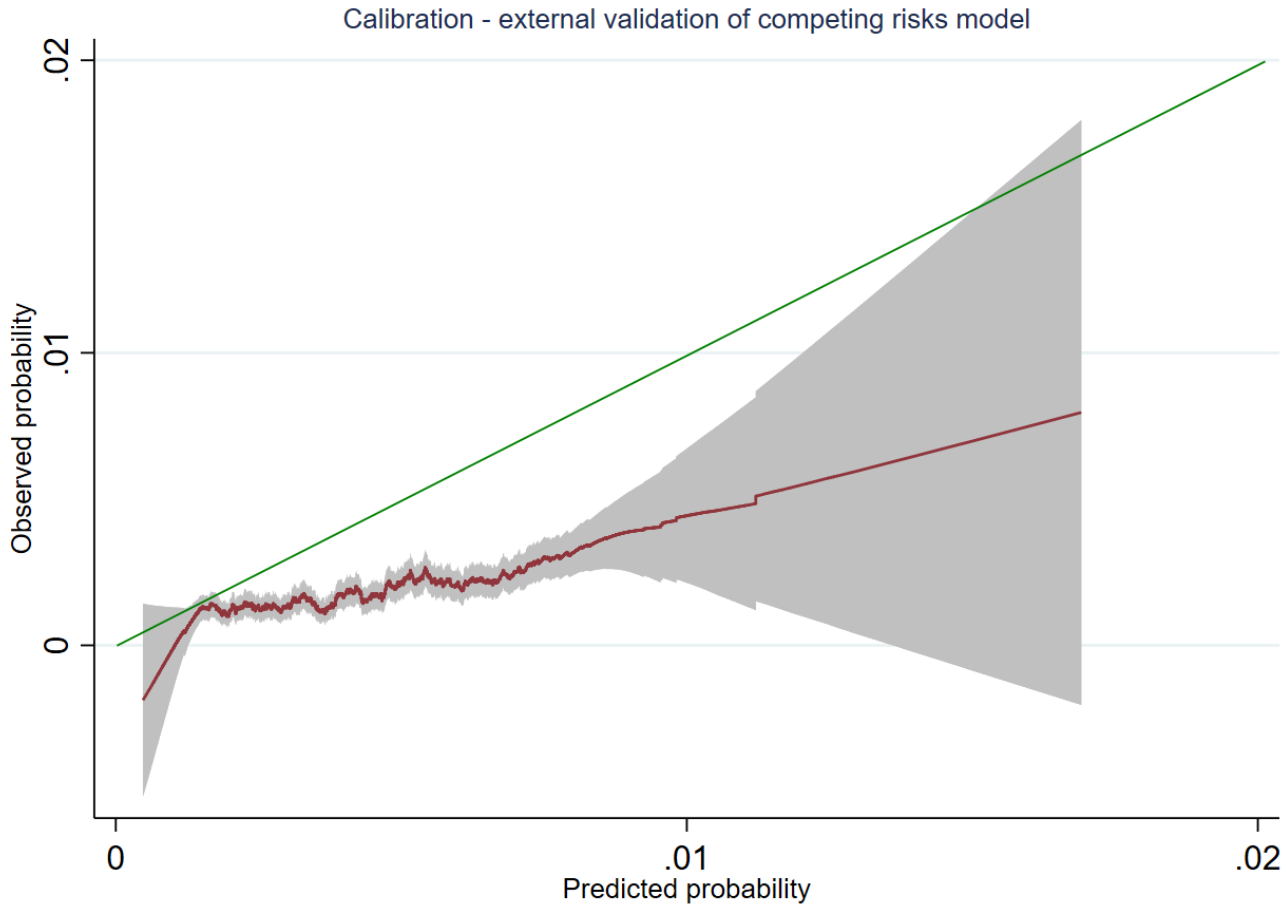

Figure 7.8. Smoothed calibration plot comparing predicted risks from the 10-year breast cancer-related mortality model, and the observed risks in the UK Biobank external evaluation cohort. Observed probabilities were obtained using pseudo-observations of the Aalen-Johansen cumulative incidence function at 10 years for breast cancer-related death. Due to the nature of pseudo-observations' calculation meaning their values are not constrained between 0 and 1, the 'observed probabilities' plotted may span zero.

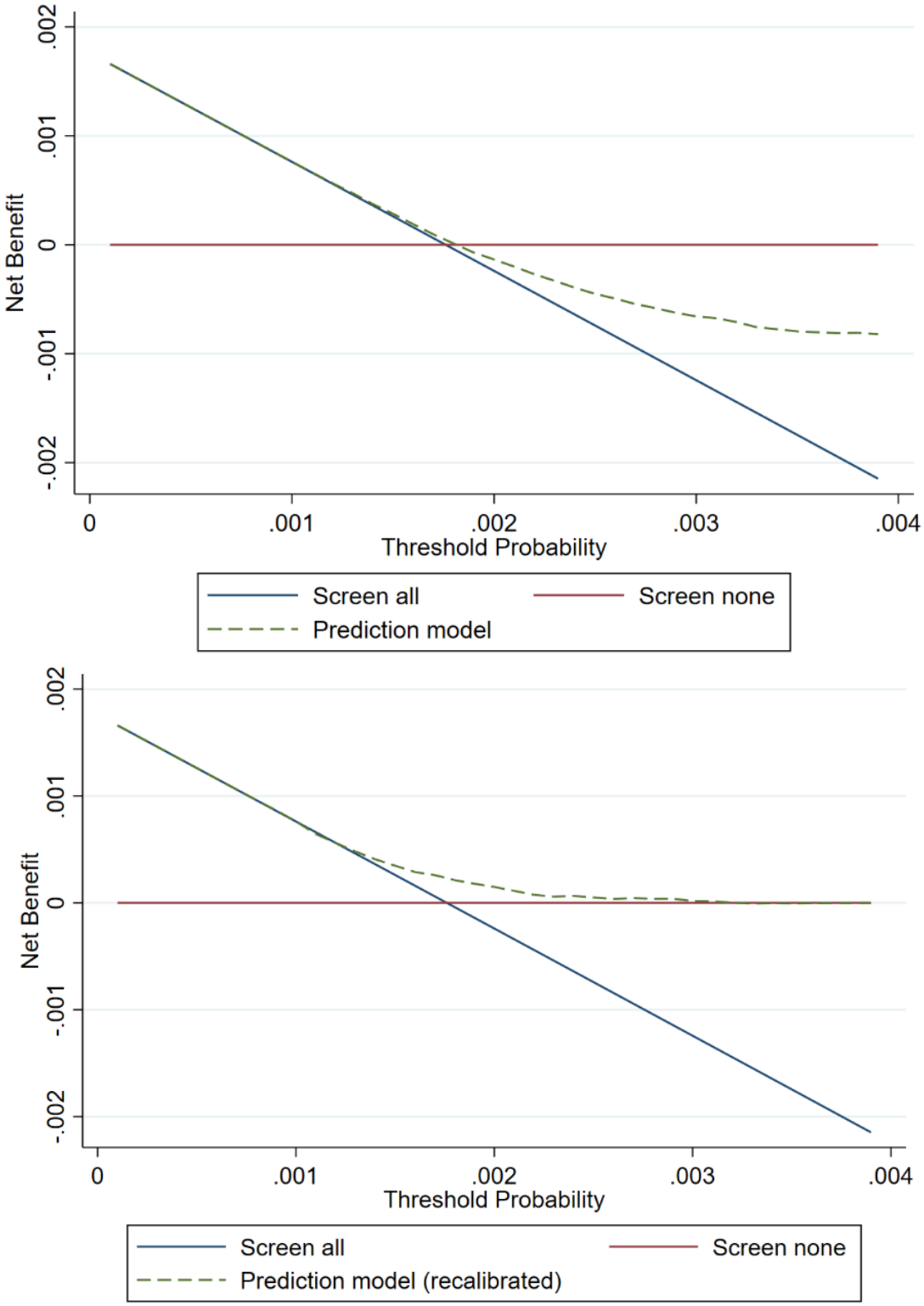

Figure 7.9. Decision curves for the competing risks model predicting 10-year breast cancer-related death risk, applied to the UK Biobank cohort. The top demonstrates the results when applying the model without any recalibration; the bottom displays the results obtained with an updated model following linear recalibration.

Integrated model development

Having evaluated the performance of 2 'phenotypic' models in UK Biobank, 4 variations of models were developed for both outcomes of interest. This was done to assess the incremental yield of incorporating additional data types alongside routinely collected phenotypic data available in primary care records. These were listed in **Chapter 6**, but repeated here for reference:

a) Model 1 ('Phenotypic') – linear predictor from QResearch model used as sole covariate
b) Model 2 ('Phenotypic + polygenic') – linear predictor from QResearch model plus the output of the genome-wide polygenic risk score (PRS) of Thompson, et al.[4]
c) Model 3 ('Phenotypic + polygenic + reproductive') – as per Model 2, but also terms for age at menarche and number of live births
d) Model 4 ('Polygenic') – only the output of the genome-wide PRS of Thompson, et al.[4]

Models for 10-year risk of incident breast cancer diagnosis

The multivariable fractional polynomial algorithm did not select non-linearities for any variables (i.e. all had powers of 1). The centred versions of each variable were used in modelling under the Cox proportional hazards framework.

The developed models are summarised as their coefficients and baseline survival functions in **Table 7.6**.

Models for 10-year risk of breast cancer death

Non-linearities were selected for the PRS output (term = 3) and age at menarche (terms = -2,-2). The linear predictor from the competing risks model and the number of live births were left linear, but centred for modelling in the competing risks framework. The developed models are summarised as their coefficients and constant term in **Table 7.7**.

	Incident breast cancer models		Breast cancer-related mortality models	
Model	Parameters	Coefficient	Parameters	Coefficient
Model 1	LP - 0.9912831921	0.3783181	LP - 1.449487315	0.6943795
(Phenotypic)	Baseline survival function	0.9661075	Constant	-6.412099
Model 2	LP - 0.9912831921	0.3442735	LP - 1.449487315	0.8036235
(Phenotypic +	PRS + 0.148714483	0.548905	PRS ^{3} – 112.8712637	0.0043341
polygenic)	(LP-0.9912831921)*(PRS + 0.148714483)	0.0095048	(LP-1.449487315)* (PRS^{3} – 112.8712637)	-0.0018433
	Baseline survival function	0.9706771	Constant	-6.526292
Model 3	LP - 0.9912831921	0.3596349	LP - 1.449487315	0.9102857
(Phenotypic + polygenic	PRS + 0.148714483	0.5489152	PRS^{3} – 112.8712637	0.0041757
+ reproductive)	(LP-0.9912831921)*(PRS + 0.148714483)	0.0090291	(LP-1.449487315)* (PRS^{3} – 112.8712637)	-0.0015643
	Menarche - 12.97421346	-0.022467	Menarche^{-2} – 0.5940711552	-0.305928
	Births - 1.830712566	-0.0550469	Menarche^{-2} * ln(menarche) – 0.1546831118	-1.772922
	Baseline survival function	0.9707544	Births – 1.830705032	-0.1067962
			Constant	-6.575861
Model 4	PRS + 0.148714483	0.5528968	PRS output^{3} – 112.8712637	0.0040156
(Polygenic)	Baseline survival function	0.970331	Constant	-6.449672

Table 7.7. Coefficients and baseline survival/constant terms for the integrated models developed and validated using the UK Biobank study cohort. LP = linear predictor from phenotypic model; PRS = output from polygenic risk score[4], "Menarche" = age at menarche in years, "Births" = number of live births. Predicted risks for the incident breast cancer models are calculable using the formula: 1 – baseline survival function$^{(\exp(\text{linear predictor}))}$. Predicted risks for the breast cancer mortality models are calculable using the formula: 1 – exp(-exp(coefficients + constant)).

Integrated model evaluation

Models for 10-year risk of incident breast cancer diagnosis

Bootstrapping estimated minimal optimism in all 4 models, with corrected calibration slopes ranging between 1.0001 (Model 1) to 1.0010 (Model 4). No coefficient shrinkage was applied due to the small degree of optimism. Optimism-corrected estimates of Harrell's C for each of Models 1 to 4 were: 0.550, 0.660, 0.661, and 0.654, respectively. **Table 7.8** summarises the overall performance of the 4 models estimated after IECV, and **Table 7.9** does so for each ethnic group. Assessment centre-specific and random effects meta-analysis pooled estimates of Harrell's C and the calibration slope for each model are presented in **Figures 7.10** and **7.11**.

The polygenic risk score-based Cox model had higher discrimination than the phenotypic-only Cox model (Harrell's C 0.654 [95% CI: 0.647-0.660] vs. 0.551 [95% CI: 0.543-0.559]), but Model 2, which incorporated both predictors had a higher discrimination point estimate than either, with a random effects meta-analysis pooled Harrell's C of 0.660 (95% CI: 0.653 to 0.666, 95% PI: 0.640 to 0.680). The confidence intervals around the Harrell's C for Models 2-4 were overlapping, however, and no model was systematically miscalibrated on summary metrics. The smoothed calibration curves showed generally good alignment for all models, with some possible overestimation of those at highest risk (Models 2 & 4 shown as exemplars in **Figure 7.13**). Explained variation was similar for Models 2-4, with overlapping confidence intervals.

Decision curve analysis showed the superiority of integrated and polygenic-only models in terms of net benefit compared to phenotypic only model (**Figure 7.13**), with the curves

for Models 2-4 being superimposed. Based on their higher discrimination, acceptable calibration, and high net benefit as per decision curve analysis, integrated and/or PRS-based models appeared to represent the best performing approaches to stratification on incident breast cancer diagnosis risks in this cohort. Using or integrating genome-wide polygenic risk information showed improvements in discrimination across ethnic groups (**Table 7.8**), with no clear effect seen on calibration (with the strong caveat of very wide confidence intervals around slope estimates, reflecting imprecision due to reduced ethnic diversity in the UK Biobank).

Performance metric	Model 1 (Phenotypic)	Model 2 (Phenotypic + PRS)
Harrell's C	0.551 (0.543-0.559) [0.522-0.580]	0.660 (0.653-0.666) [0.640-0.680]
Calibration slope	1.042 (0.850-1.234) [0.322-1.762]	0.995 (0.951-1.039) [0.851-1.139]
R^2	0.018 (0.012-0.024) [0.00-0.041]	0.170 (0.157-0.183) [0.129-0.211]
Performance metric	**Model 3 (Phenotypic + PRS + reproductive)**	**Model 4 (PRS only)**
Harrell's C	0.661 (0.654-0.667) [0.640-0.681]	0.654 (0.647-0.660) [0.632-0.675]
Calibration slope	0.994 (0.949-1.040) [0.845-1.144]	0.994 (0.944-1.045) [0.821-1.168]
R^2	0.172 (0.159-0.185) [0.129-0.215]	0.158 (0.145-0.171) [0.114-0.202]

Table 7.8. Final performance estimates for 4 model variations predicting 10-year incidence breast cancer diagnosis risk, fit to the UK Biobank data – these are random effects meta-analysis pooled estimates obtained from internal-external cross-validation. Point estimates are reported with (95% confidence intervals) and [95% prediction intervals]. PRS = genome-wide polygenic risk score.

Ethnic group	Performance metric	Model 1 (Phenotypic)	Model 2 (Phenotypic + PRS)	Model 3 (Phenotypic + PRS + reproductive)	Model 4 (PRS only)
White	Harrell's C	0.548 (0.541-0.554)	0.662 (0.656-0.668)	0.662 (0.656-0.668)	0.656 (0.650-0.662)
	Calibration slope	0.964 (0.841-1.087)	1.014 (0.976-1.052)	1.012 (0.975-1.050)	1.024 (0.985-1.064)
	R^2	0.017 (0.013-0.022)	0.175 (0.159-0.181)	0.177 (0.166-0.187)	0.165 (0.154-0.176)
South Asian	Harrell's C	0.588 (0.522-0.642)	0.633 (0.577-0.689)	0.636 (0.579-0.692)	0.614 (0.558-0.671)
	Calibration slope	1.514 (0.477-2.550)	0.861 (0.511-1.210)	0.872 (0.528-1.215)	0.771 (0.407-1.135)
	R^2	0.052 (0.006-0.131)	0.130 (0.050-0.227)	0.140 (0.056-0.239)	0.099 (0.030-0.193)
Other Asian	Harrell's C	0.616 (0.482-0.750)	0.654 (0.508-0.800)	0.657 (0.513-0.800)	0.643 (0.492-0.794)
	Calibration slope	1.809 (-0.617-4.235)	0.927 (0.138-1.716)	0.926 (0.140-1.713)	0.810 (0.005-1.614)
	R^2	0.070 (0.001-0.294)	0.170 (0.007-0.406)	0.168 (0.007-0.404)	0.123 (0.002-0.358)
Black	Harrell's C	0.578 (0.516-0.640)	0.630 (0.574-0.687)	0.640 (0.585-0.696)	0.613 (0.556-0.671)
	Calibration slope	1.625 (0.442-2.807)	0.725 (0.391-1.058)	0.770 (0.440-1.099)	0.657 (0.313-1.000)
	R^2	0.044 (0.003-0.124)	0.109 (0.034-0.208)	0.125 (0.045-0.226)	0.088 (0.021-0.182)
Chinese	Harrell's C	0.422 (0.309-0.534)	0.614 (0.482-0.745)	0.613 (0.477-0.750)	0.637 (0.504-0.769)
	Calibration slope	-1.58 (-4.542-1.367)	0.790 (-0.066-1.647)	0.832 (-0.015-1.678)	1.004 (0.108-1.900)
	R^2	0.027 (0.001-0.217)	0.099 (0.001-0.326)	0.112 (0.001-0.339)	0.143 (0.004-0.373)
Mixed race/other	Harrell's C	0.529 (0.474-0.583)	0.602 (0.549-0.655)	0.601 (0.548-0.655)	0.599 (0.546-0.653)

	Calibration slope	0.534 (-0.488-1.557)	0.609 (0.307-0.910)	0.590 (0.291-0.889)	0.610 (0.298-0.922)
	R^2	0.005 (0.001-0.046)	0.078 (0.020-0.160)	0.075 (0.019-0.156)	0.075 (0.019-0.156)
Not recorded	Harrell's C	0.521 (0.362-0.679)	0.633 (0.503-0.764)	0.648 (0.520-0.776)	0.628 (0.495-0.759)
	Calibration slope	0.968 (-1.943-3.880)	0.660 (-0.173-1.493)	0.710 (-0.112-1.532)	0.615 (-0.240-1.469)
	R^2	0.015 (0.001-0.217)	0.089 (0.001-0.334)	0.107 (0.001-1.356)	0.078 (0.001-0.320)

Table 7.9. Ethnic group-specific performance metrics (with 95% confidence intervals) for the 4 integrated models fit and evaluated in UK Biobank. PRS = genome-wide polygenic risk score.

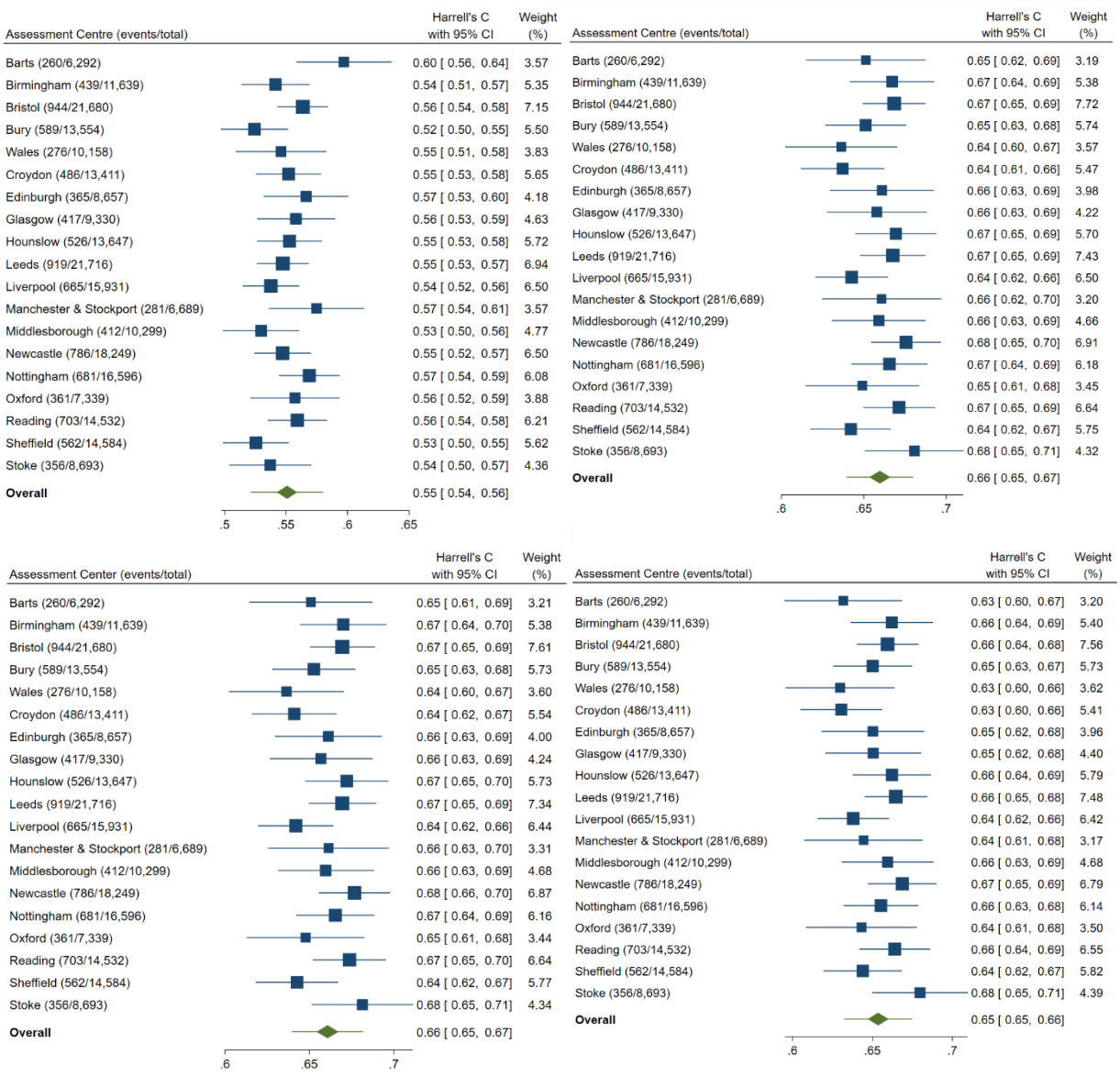

Figure 7.10. Regional estimates of Harrell's C for the 4 integrated models predicting 10-year incident breast cancer diagnosis risk, with pooled meta-estimates. Model 1 = top left, Model 2 =top right, Model 3 = bottom left, Model 4 = bottom right.

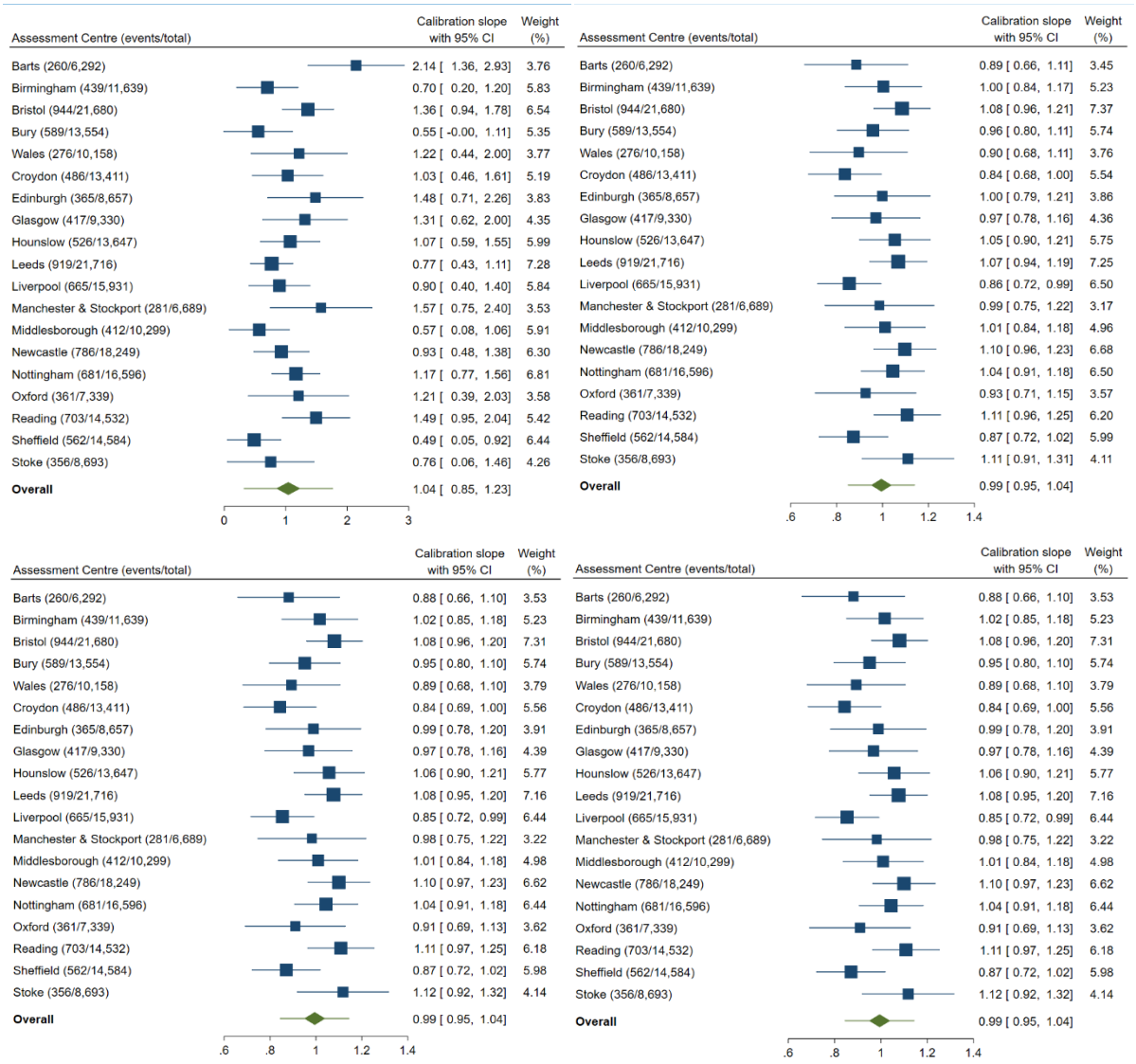

Figure 7.11. Regional estimates of the calibration slope for the 4 integrated models predicting 10-year incident breast cancer diagnosis risk, with pooled meta-estimates. Model 1 = top left, Model 2 =top right, Model 3 = bottom left, Model 4 = bottom right.

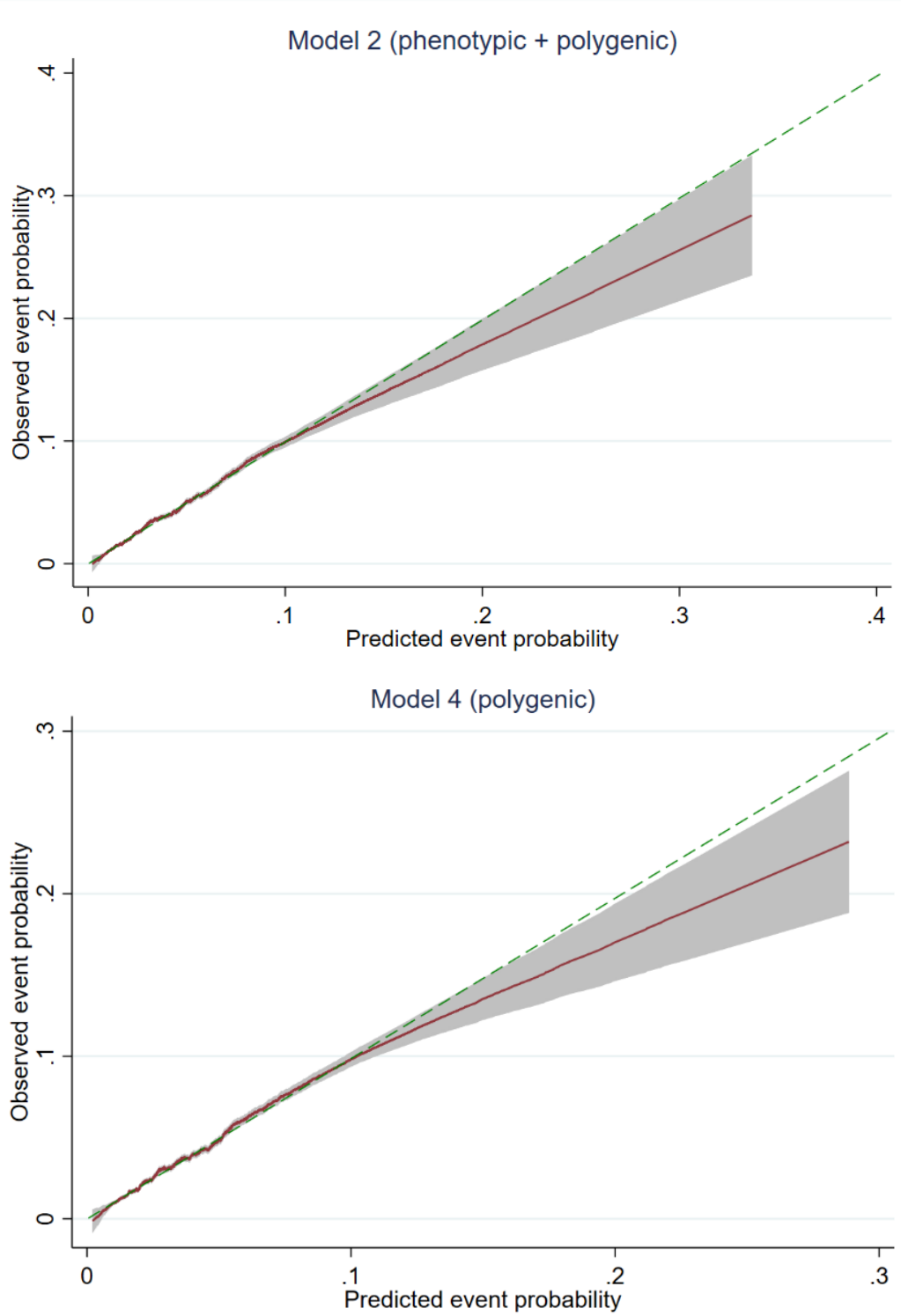

Figure 7.12. Smoothed calibration plots for 2 models predicting 10-year incident breast cancer diagnosis risk, fit using data from UK Biobank. Predicted probabilities for both models are those calculated during the internal-external cross-validation process.

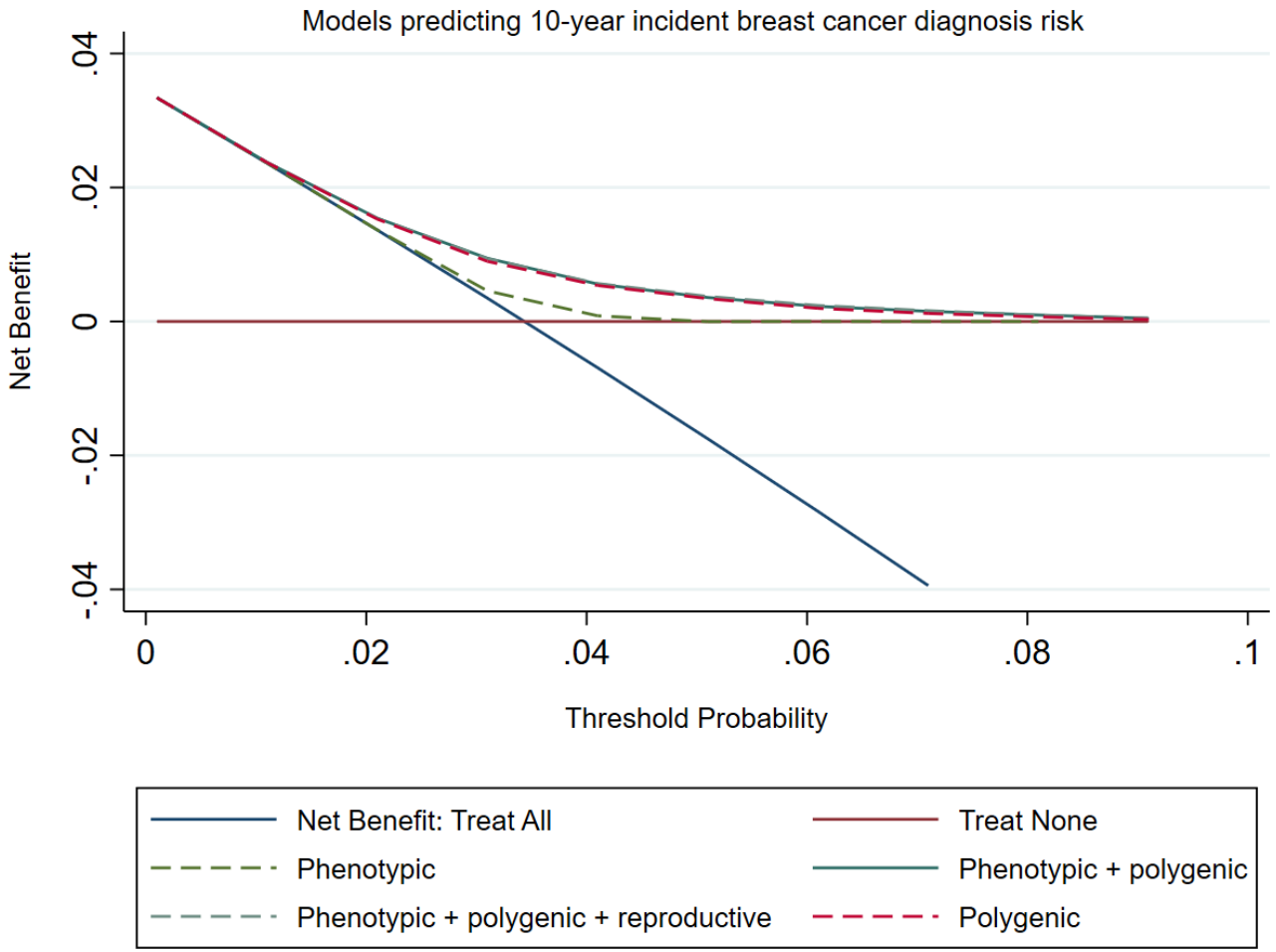

Figure 7.13. Decision curves comparing the clinical utility of the 4 models developed using UK Biobank data to predict the 10-year risk of incident breast cancer diagnosis.

Models for 10-year risk of breast cancer-related death

Bootstrapping estimated a small degree of optimism in all 4 models, with corrected calibration slopes ranging between 0.998 (Model 4) to 1.013 (Model 3). Optimism-corrected estimates of Harrell's C for each of Models 1 to 4 were: 0.582 (95% CI: 0.554 to 0.616), 0.650 (95% CI: 0.620 to 0.686), 0.637 (95% CI: 0.593 to 0.677) and 0.635 (95% CI: 0.610 to 0.663), respectively. **Table 7.10** summarises overall performance metrics for the 4 models estimated after IECV.

Due to low event counts in different groups, ethnicity-specific performance metrics were not calculated. Assessment centre-specific and random effects meta-analysis pooled estimates of Harrell's C are presented in **Figure 7.14**.

Similar to the incident breast cancer prediction models, integration of genome-wide polygenic risk had an incremental benefit on discrimination, with a phenotypic-only model's random effects meta-analysis pooled Harrell's C of 0.583 (95% CI: 0.546-0.621) compared to 0.657 (95% CI: 0.626-0.688) for Model 2. Integration of reproductive factor information did not have any clear impact on discrimination capability above the other 2 sources of information. Although the PRS was developed for incident breast cancer, its conversion to use in the current risk trajectory suggests that this genome-wide summation of risk may be relevant to predicting not just 'any' breast cancer, but life-threatening ones (random effects meta-analysis pooled Harrell's C of Model 4: 0.646, 95% CI: 0.619-0.672). Smoothed calibration curves for all 4 models showed appropriate alignment of observed and predicted risks (Model 2 and Model 4 displayed in **Figure 7.15**), with uncertainty around calibration in those at the very highest predicted risks (e.g. >0.005 probability) due to very low numbers of people with such predictions generated.

Decision curve analysis of these 4 models (**Figure 7.16**) showed that whilst Model 4 (genome-wide PRS) had improved net benefit to a phenotypic-only model (Model 1), the integrated models had superior clinical utility to both, across most of the threshold probability range. The decision curves for Models 2 & 3 were effectively superimposed across the threshold probabilities examined. Overall, models that integrated both phenotypic and genotypic sources of information appeared to offer the best performing approaches to stratifying screening-age women in terms breast cancer-related death risks in this cohort.

Performance metric	Model 1 (Phenotypic)	Model 2 (Phenotypic + PRS)
Harrell's C	0.583 (0.546-0.621) [0.449-0.718]	0.657 (0.626-0.688) [0.551-0.763]
Calibration slope	1.000 (0.983-1.016) [0.950-1.050]	1.000 (0.984-1.017) [0.949-1.051]
Performance metric	**Model 3 (Phenotypic + PRS + reproductive)**	**Model 4 (PRS only)**
Harrell's C	0.649 (0.617-0.682) [0.539-0.760]	0.646 (0.619-0.672) [0.559-0.733]
Calibration slope	1.346 (0.574-2.118) [-2.134-4.826]	0.998 (0.983-1.103) [0.954-1.042]

Table 7.10. Final performance estimates for 4 models predicting 10-year breast cancer mortality risk, fit to the UK Biobank data – these are random effects meta-analysis pooled estimates obtained from internal-external cross-validation. Point estimates are reported with (95% confidence intervals) and [95% prediction intervals]. PRS = genome-wide polygenic risk score.

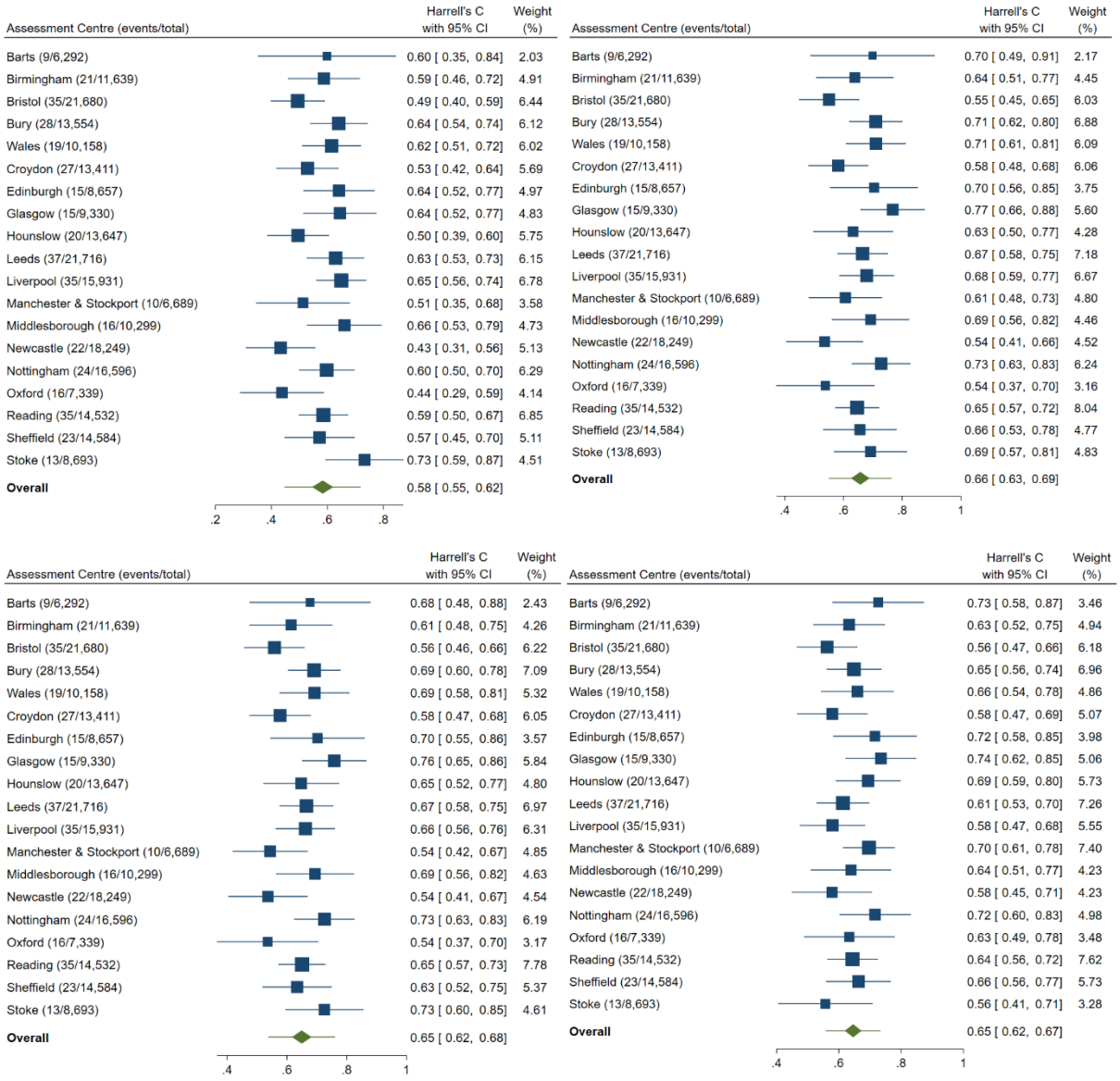

Figure 7.14. Regional estimates of Harrell's C for the 4 integrated models predicting 10-year breast cancer mortality risk, with pooled meta-estimates. Model 1 = top left, Model 2 =top right, Model 3 = bottom left, Model 4 = bottom right.

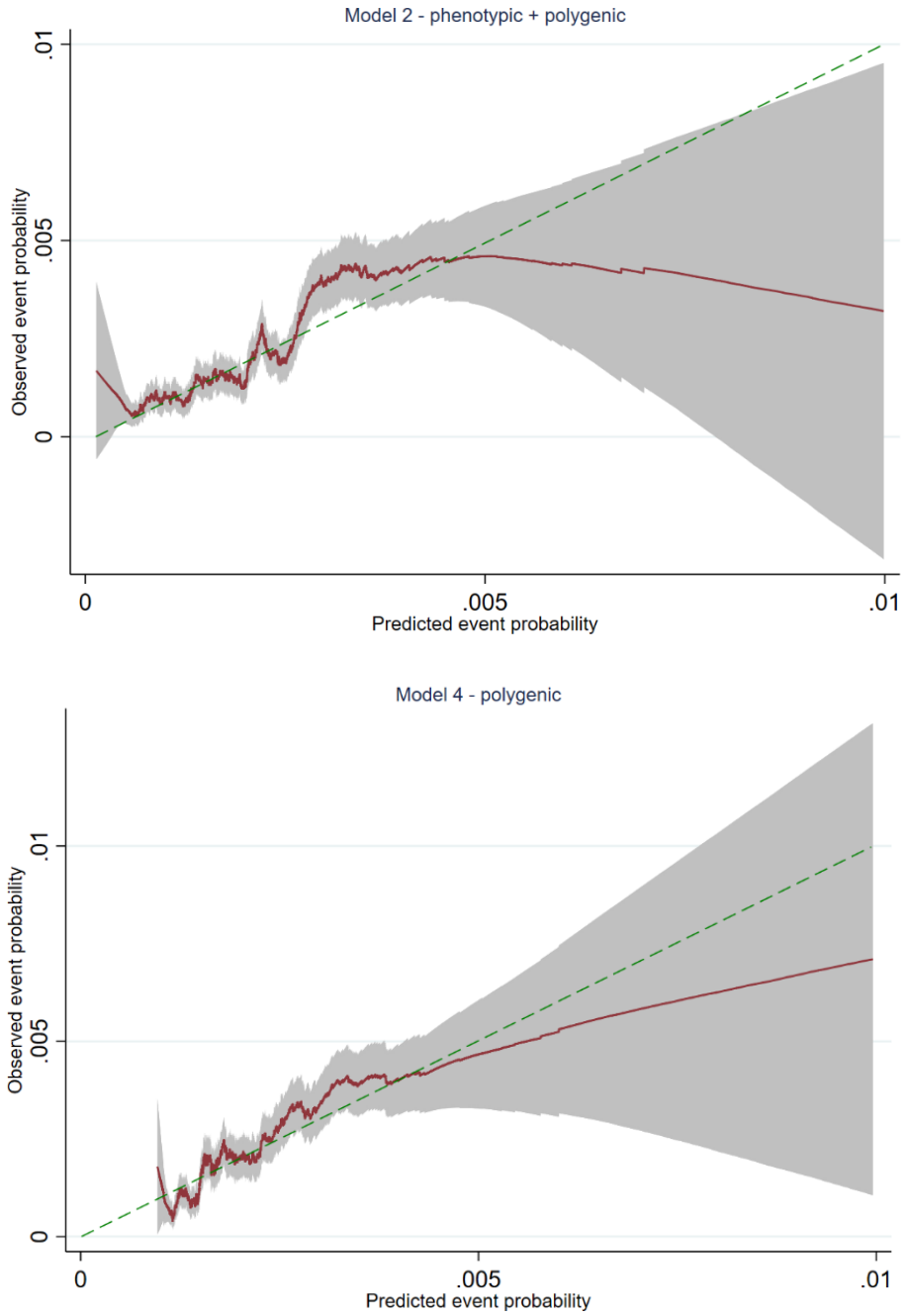

Figure 7.15. Smoothed calibration plots for 2 models predicting 10-year breast cancer-related mortality risk, fit using data from UK Biobank. Predicted probabilities for both models are those calculated from internal-external cross-validation process.

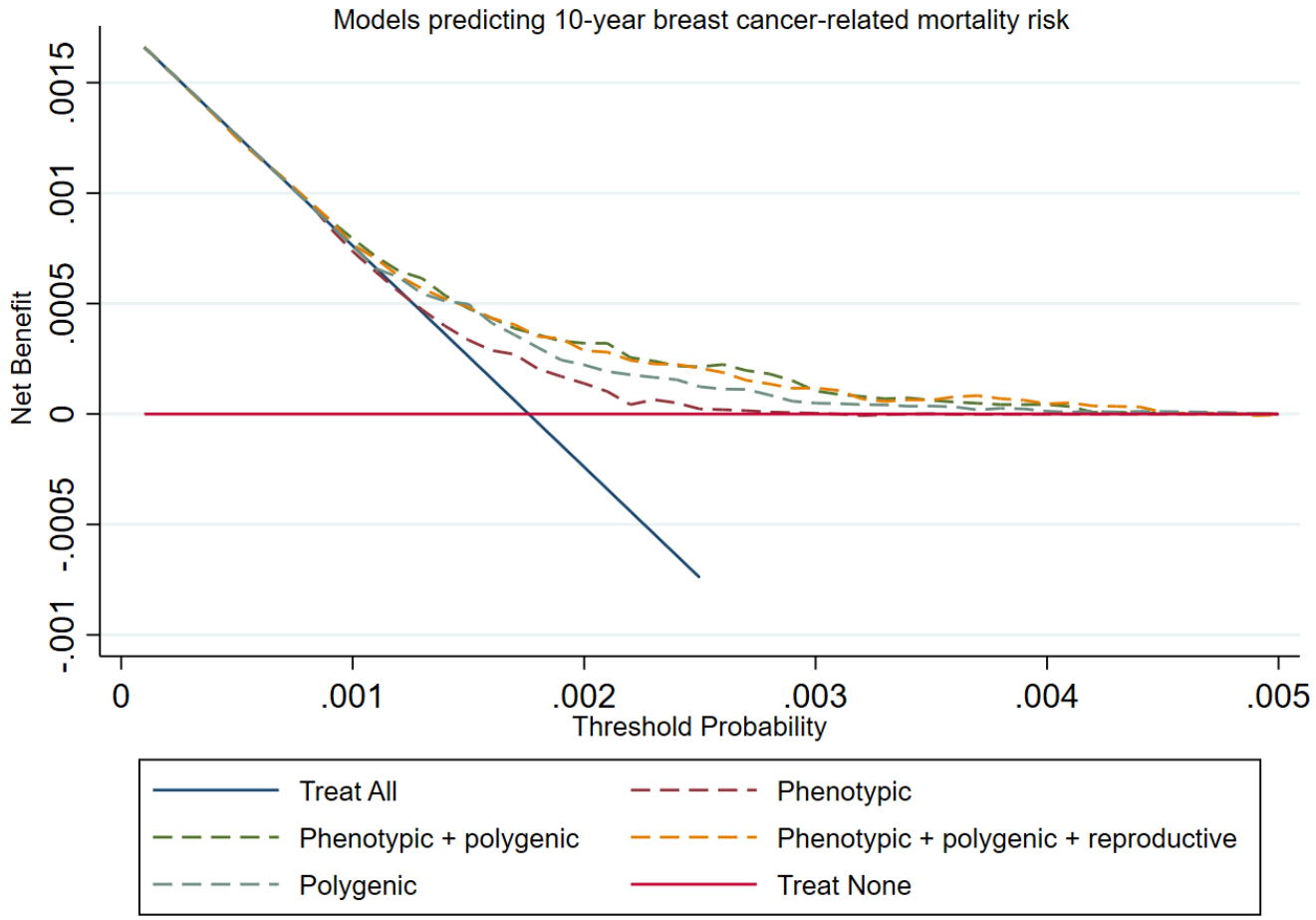

Figure 7.16. Decision curves comparing the clinical utility of the 4 models developed using UK Biobank data to predict the 10-year risk of breast cancer-related death.

Discussion

The current NHS Breast Screening Programme invites women aged 50-70 years of age, and the US Preventive Task Force guidelines recommend screening women aged 50-74 years of age, with shared decision making for women aged 40-49years[5]. The efficacy and cost-effectiveness of screening women in the general population outside these age ranges is not yet fully determined[6-8], and it is perhaps likeliest that any risk-stratified population breast screening programme would invite women of similar age ranges to these. These age ranges are important to consider for two reasons: first, the concept of targeted validation and second, the role of sample heterogeneity on model performance:

1) Targeted validation refers to the evaluation of a clinical prediction model for a specific task, i.e. the model's intended use and its target population[9]. The phenotypic models developed using QResearch in **Chapters 3 & 5** used data for women aged 20-90 years, thereby representing divergence between the sample used for model development, the sample from UK Biobank used in this chapter, and the population in which the model(s) could ultimately be implemented. The results in this chapter supplement the age group-specific performance metrics reported in earlier chapters such as the similar age-group discrimination results across cohorts. More importantly, they further the understanding of model performance (and how well models may 'need' to perform) in screening age women to inform risk-stratified screening strategies. Of note, whilst the UK Biobank may meet criteria for targeted validation with respect to its age profile, it has several limitations which may reduce its appropriateness for use as a targeted validation for the UK's screening-eligible population (see below).

2) Age is the strongest predictor for breast cancer, therefore models may be expected to exhibit reduced discrimination in populations with a narrower age range, due to restricted heterogeneity within the predominant predictor. Clinical prediction models for breast cancer that use only phenotypic predictors tend to have modest discrimination performance when restricted to screening age women[2,3,10] and higher discrimination when tested in populations including women outside screening age eligibility[11] – the models developed in **Chapters 3 & 5** are no different. Therefore, the key results in this chapter are the assessment of incremental effects on model performance on integrating additional data modalities that were not available in the derivation data but could be obtainable from women invited to a national risk-based screening programme.

The phenotypic breast cancer incidence Cox model developed using QResearch (**Chapter 3**) demonstrated poor performance in this selected, less heterogeneous UK Biobank cohort. The results from integrated modelling reinforce the notion that use of additional data (i.e. genome-wide polygenic risk) is likely to be needed to perform accurate assessment of incident breast cancer diagnosis risks with sufficient clinical utility in this target population. Interestingly, decision curve analysis suggested that the majority of prognostication ability in this scenario was captured by the PRS, indeed, Model 2 had an increment in Harrell's C of 0.006 over Model 4. Previous studies of integrated modelling (phenotypic, reproductive, mammographic, and PRSs with far fewer SNPs) show incremental improvements in discrimination with doing so, but generally do not report decision analysis curves to compare clinical utility[3,10]. This chapter extends this prior

work, which typically use smaller PRSs (e.g. 18 gene or 313 SNP panels[12]) to using a genome-wide PRS comprising over 6 million SNPs.

In contrast, the breast cancer mortality model had slightly better discrimination on application to the UK Biobank cohort, and the integrated modelling suggested that the best approach (in terms of net benefit) for risk assessment in this scenario was to combine phenotypic with polygenic information. The combination of phenotypic and genome-wide polygenic risk (for a related, but separate trait) had higher clinical utility than either data source alone, and the highest net benefit overall, suggesting that Model 4 combined orthogonal but complementary information relevant to risk. The work in this chapter provides the first validation of the Thompson, et al. genome-wide PRS[4] for a proxy trait, but also extends the approach of **Chapter 5** in directly modelling breast cancer mortality risk into an integrated data setting. Unlike previously observed null or inverse associations between risk of incident breast cancer predicted by some phenotypic models and risk of mortality after diagnosis[13], the genome-wide PRS-including models (Models 2-4) appear to capture information about proclivity to life-threatening cancers. Ability to predict cancer mortality risk has been proposed to be more informative to stratified screening, for example in the setting of prostate cancer[14].

Data from UK Biobank and its linked datasets present opportunities and limitations. The breadth of data collection and individual participant-level linkages from UK Biobank across Hospital Episodes Statistics, primary care data and national registries permitted robust ascertainment of predictor values and outcomes and facilitated integrated phenotypic and genetic modelling. Further, the ability of UK Biobank researchers to return results permitted the direct use of a novel genome-wide PRS's output[4] as a predictor in the current work. However, UK Biobank is a cohort with evidence of selection bias – UK Biobank participants (5.5% of the 9.2million total invited) are for

example, more likely to be female, live in more affluent areas, have fewer self-reported comorbidities, and have lower all-cause mortality risks than the wider sample population[15]. These phenomena have predicated concern regarding selection and collider bias in epidemiological studies[16,17], but these may also have relevance to prediction modelling analyses. First, the models fit and cross-validated in UK Biobank may be miscalibrated on application to population-representative data (a truer targeted validation), or when used in populations undergoing screening. Whilst the age group-stratified breast cancer incidence rates were similar between UK Biobank and the national-level statistics (**Table 7.2**), they were lower for breast cancer mortality (**Table 7.3**), and given UK Biobank's healthy volunteer bias, the predictor distributions may be different to those expected in the UK sub-population eligible for the NHS BSP, or a future risk-stratified version thereof. Whilst selection or collider bias can distort the magnitude and/or direction of estimands in an observational or causal inference setting[16,17], and though coefficients in these clinical prediction models do not have a causal interpretation, skewed/inappropriate predictor-outcome associations could be estimated in non-representative data. This can impact face validity of derived models, or their performance on application to external datasets. Finally, the reduced ethnic diversity of UK Biobank is another manifestation of non-representativeness – this combined with the reduced breast cancer mortality rate precluded an ethnic group-specific performance assessment of the mortality models.

One approach to counteract this selection bias could be the use of sampling weights during model fitting. A recent pre-print study considered the restrictions on the UK Biobank-eligible population in terms of age range and geographical location (22 assessment centres), and used UK Census microdata for the eligible sampling populations to estimate inverse probability weights for UK Biobank participants (i.e. inversely

proportional to the likelihood of their participation in the cohort) using probit regression modelling[18]. Whilst this pre-publication report suggests that bias in estimated coefficients from selected models may be reduced by up to 78% with use of these sampling weights[18], they were not available from the UK Biobank research access platform at time of this chapter's work being conducted. It would be of interest to later re-perform this chapter's analyses using these sampling weights and compare the performance of unweighted and weighted integrated models in a distinct, representative population.

In conclusion, integrating the output from a 6 million SNP genome-wide PRS into phenotypic clinical prediction models using routinely collected clinical data improved the performance of models for incident breast cancer diagnosis and breast cancer mortality in a cohort of women in the UK Biobank. This was not clearly improved upon by further consideration of reproductive factor information. Although the use of data from women aged 40-69 years provided a partial targeted validation, selection bias inherent to the UK Biobank means that the transportability of results to all screening age women in the UK should be interpreted carefully. In these risk prediction scenarios, with the routinely available predictors selected for the models, and considering the relatively narrow range of age of screening eligibility under the NHS BSP, these results suggest that using phenotypic data alone may not be sufficient for accurate risk assessment in screening age women if risk-based screening strategies were based on breast cancer diagnosis risks. However, in the breast cancer mortality prediction scenario, combining phenotypic and genotypic information offered complementary information and represented the approach associated with highest clinical utility in this selected cohort. Whether the clinical impacts and cost-effectiveness of such integrated models are superior to models using phenotypic data alone is worthy of evaluation. Whether there is additional scope for improved model

performance with integration of mammographic features (e.g. breast density) is another avenue for future study.

Chapter references

1. Booth S, Riley RD, Ensor J, et al. Temporal recalibration for improving prognostic model development and risk predictions in settings where survival is improving over time. *Int J Epidemiol* 2020; **49**(4): 1316-25.

2. Brentnall AR, Cuzick J, Buist DSM, et al. Long-term Accuracy of Breast Cancer Risk Assessment Combining Classic Risk Factors and Breast Density. *JAMA Oncol* 2018; **4**(9): e180174.

3. 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**(4): 476-82.

4. Thompson DJ, Wells D, Selzam S, et al. UK Biobank release and systematic evaluation of optimised polygenic risk scores for 53 diseases and quantitative traits. *medRxiv* 2022: https://www.medrxiv.org/content/10.1101/2022.06.16.22276246v2.

5. Siu AL, US Preventive Task Force. Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement. *Ann Intern Med* 2016; **164**(4): 279-96.

6. Schousboe JT, Sprague BL, Abraham L, et al. Cost-Effectiveness of Screening Mammography Beyond Age 75 Years : A Cost-Effectiveness Analysis. *Ann Intern Med* 2022; **175**(1): 11-9.

7. Duffy SW, Sasieni P, Olsen AH, et al. Modelling the likely effect of the increase of the upper age limit from 70 to 73 for breast screening in the UK National Programme. *Stat Methods Med Res* 2010; **19**(5): 547-55.

8. Duffy SW, Vulkan D, Cuckle H, et al. Effect of mammographic screening from age 40 years on breast cancer mortality (UK Age trial): final results of a randomised, controlled trial. *Lancet Oncol* 2020; **21**(9): 1165-72.

9. Sperrin M, Riley RD, Collins GS, et al. Targeted validation: validating clinical prediction models in their intended population and setting. *Diagn Progn Res* 2022; **6**(1): 24.

10. Evans DGR, van Veen EM, Harkness EF, et al. Breast cancer risk stratification in women of screening age: Incremental effects of adding mammographic density, polygenic risk, and a gene panel. *Genet Med* 2022; **24**(7): 1485-94.

11. Terry MB, Liao Y, Whittemore AS, et al. 10-year performance of four models of breast cancer risk: a validation study. *Lancet Oncol* 2019; **20**(4): 504-17.

12. 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**(1): 21-34.

13. 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**(8): e0160966.

14. Vickers AJ, Sud A, Bernstein J, et al. Polygenic risk scores to stratify cancer screening should predict mortality not incidence. *NPJ Precis Oncol* 2022; **6**(1): 32.

15. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* 2017; **186**(9): 1026-34.

16. Griffith GJ, Morris TT, Tudball MJ, et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat Commun* 2020; **11**(1): 5749.

17. Munafo MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. *Int J Epidemiol* 2018; **47**(1): 226-35.

18. van Alten S, Domingue BW, Galama T, Marees AT. Reweighting the UK Biobank to reflect its underlying sampling population substantially reduces pervasive selection bias due to volunteering. *MedRxiv* 2022.

https://www.medrxiv.org/content/10.1101/2022.05.16.22275048v1.full

Chapter Eight

Conclusions and avenues for future research

Summary

This final chapter briefly recapitulates the motivations for the thesis, then provides a summary of insights and key results generated during it and their contributions to the existing literature. By considering the work undertaken and its results within the context of the 'leaky pipeline'[1] of prediction model research from study inception to model implementation, it highlights points of interest, considers methodological and ethical questions that have emerged during the analyses, the limitations of the work undertaken, and sets out a framework for future research.

Thesis motivation

The underlying motivation of this thesis was to contribute to the ability to make better public health and clinical decisions that reduce deaths from breast cancer. The initial conception of the thesis was centered around risk-based screening. Risk-based screening seeks to tailor breast cancer screening strategy to the risk profile of individual women, but similar paradigms are applicable to other approaches to reduce breast cancer mortality, e.g. breast cancer prevention (e.g. how can we identify women most suitable for chemoprevention?) and the management of breast cancers after diagnosis (e.g. can we

stratify women into different forms of follow-up?). All of these are rooted in a base dependency of clinical prediction models that can accurately and reliably estimate individual women's risks. However, these areas are severely limited by the following:

1) Almost all existing models are at high risk of bias and cannot be recommended for use[2,3]
2) The methodological quality of clinical prediction model development and evaluation is generally poor
3) Predicted risks of developing breast cancer and dying from breast cancer are poorly correlated, but the risk-based screening/prevention field thus far has focused only on models that predict incidence

Contributions to the literature

This thesis sought to develop clinical prediction models that could be useful for risk-based public health strategies for breast cancer early detection/prevention or clinical post-diagnosis management strategies, compare the benefits and limitations of different algorithmic approaches, and evaluate their performance.

For two of the endpoints, i.e. risk of incident breast cancer and risk of mortality following diagnosis, the act of developing prediction models is not novel[2,3]. As stated in previous chapters and appraised in recent systematic reviews, there are multiple alternative models for these outcomes but very few (if any) are reliable[4-8]. The contributions of this thesis to the field are manifest in:

1) The development and evaluation of clinical prediction models in a framework that targeted fair comparisons in the context of growing concerns of poorly transparent reporting and 'spin'[4-7]
2) Application of a validation strategy that sought to maximise understanding of model performance and heterogeneity if applied to new settings[9]
3) The first work to model the combined risk of developing and dying from breast cancer in the general female population without breast cancer at baseline
4) The assessment of integrating the largest ever developed genome-wide polygenic risk score to models that use routine clinical data (with a proxy trait validation)

For models predicting breast cancer incidence, two systematic reviews within the last 5 years have cautioned against recommending any single one for use in stratified screening[3,10], although neither of these appraised the QCancer-10year Breast[11] model which is currently in clinical use in the UK. This is due to the search strings used for the literature searches.

However, as discussed in **Chapters 1 & 5**, this thesis posits that structuring screening or prevention strategy around incident breast cancer diagnosis risks is neither the only nor necessarily the optimal approach to further reduce mortality from this common cancer[12]. The main contribution of this thesis to the existing body of knowledge is the consideration of breast cancer risks from different perspectives – specifically the work in **Chapter 5** which represents the first study to develop a model to directly predict risk of life-threatening breast cancers in the general female population without breast cancer at baseline. Since the work for this thesis began, a new model was reported that estimates the 6-year risk that a woman undergoing screening will be diagnosed with an 'advanced' breast cancer, defined as stage II or higher[13]. Developed using data from over 930,000

women in the Breast Cancer Surveillance Consortium, the final model included predictors such as obesity, high breast density and proliferative breast disease with atypia. Whilst a partial evolution from simply modelling incident diagnosis risks, this study only included women already attending a screening program, only used weak calibration metrics, had no assessment of clinical utility, and did not explore the correlation of outputs with breast cancer death[13].

The work in this thesis that developed models to provide reliable risk assessment for any women with breast cancer could (pending further evaluation) have clinical utility in the post-diagnostic pathway. Following review of the one model deemed to be at low risk of bias in the systematic review of 922 models by Hueting, et al. for this outcome, it is likely that no reliable tool exists to provide this[2].

The UK Biobank work package (**Chapters 6 & 7**), whilst not able to provide a fully meaningful external validation due to selection bias and sampling design limiting its applicability to the target population[14-16], did provide an opportunity to estimate the incremental effects of combining genetic and phenotypic information for risk assessment should there be wider availability of genetic sequencing information in the future. Other breast cancer incidence models have integrated polygenic and phenotypic information[17-20], but this has been with smaller gene panels in contrast to the genome-wide PRS of Thompson, et al. which assimilates the effects of genetic loci orders of magnitude larger in number[21]. The improvement in discrimination was relatively modest, and given the scale of the PRS, it is perhaps unlikely that further gains can be attained in predictive accuracy with the addition of further variables. Many biological processes are functions of causal factors but also random 'noise' that is neither measurable nor predictable, so for most (if not all) prognostic modelling tasks, there will probably be a ceiling of performance that can be attained.

This work in this thesis alone cannot support the clinical implementation of any of the models developed. Evidence to support model implementation in the target contexts (e.g. risk-based population screening strategies) needs to be multifaceted, including triangulation of information from further evaluation studies such as external/local validation, decision analysis modelling with consideration of cost-effectiveness, and qualitative research to understand perspectives of target end-users that may present barriers to utilisation.

The integration of clinical prediction models into clinical workflows

During this thesis, initially unanticipated methodological considerations emerged such as the role of using protected characteristics as predictor variables, as have methodological developments that build on existing concepts such as the value of model validation and assessing model stability. Rather than discussing the limitations of this thesis and avenues for future research introspectively, this chapter now considers a conceptual framework from model development to implementation into clinical workflows and summarises the thesis in this context. This is adapted from the pipeline as described by van Royen, et al.[1] and integrates or expands upon additional aspects where relevant.

When is a model appropriate for use?

First, we consider when a model is appropriate for use. Having defined a target population with a clearly articulated clinical need, clinical prediction model development should

integrate predictors that are readily available in the context of its intended use, and modelling should be undertaken in an appropriately representative data sample with accurate measurement of the outcome[1,15].

In this thesis, three outcomes were modelled – the risk of incident breast cancer diagnosis, the combined risk of developing and then dying from breast cancer, and the risk of mortality after a breast cancer diagnosis is made. Earlier chapters have sought to state that each of these outcomes has relevance to clinical practice and public health policy. Further, **Chapters 2-5** discussed how these models were developed using routinely collected clinical data within primary care data systems – this information is available at the 'point of care', and so aligns with the likeliest implementation environment, which could be as a software plug-in to electronic health records systems. Lastly, the ascertainment of outcomes was based on (depending on outcome) structured data within data assets linked at the individual-level, which may be as robust as possible in this scenario.

Once a model is developed, its evaluation should consider multiple facets of performance. This is because prediction tools have the potential to cause harm, and lack of validation (or a poorly performed validation analysis) limits implementation due to poor confidence in the model. Discrimination, calibration, and net benefit each contribute to understanding how well a model distinguishes between events and non-events, how aligned predicted and observed risks are, and whether a model could inform better decision making, respectively. How and in whom these are estimated affects interpretation – an extreme example is the validation of a model in a setting that is very different to its intended use but in a 'convenience dataset', which cannot be informative[15]. Even when using a dataset that reflects the target population, estimating any of these metrics in the 'overall' study population (e.g. validation sub-dataset) may not realise the full scope of variation in clinically relevant sub-groups, such as people of different ages, ethnicities, deprivation or

geography [22]. Prediction models generally estimate the 'average' magnitude and direction of associations between sets of predictors and an outcome with uncertainty. These do not necessarily transport into all settings and don't necessarily translate into accurate risk assessment at the individual level[23], but the implicit assertion by papers reporting only 'overall' metrics is that model predictions are equally reliable and robust for all individuals. Whilst this may not be an issue when a model is intended to be used in a narrowly defined population, it becomes increasingly important when models are designed to be deployed on larger target populations.

In **Chapters 3 & 5**, models were developed using data representing the general female population. In **Chapter 4**, this was in the unselected female population diagnosed with invasive breast cancer. All of these cohorts are heterogeneous in their composition, e.g. age, ethnicity and the clinicopathological characteristics of the diagnosed tumours. Therefore, assessing performance heterogeneity is important to consider. In this thesis, model performance was assessed overall and in different age and ethnic groups. For the prognostic models (**Chapter 4**), this was also performed for different tumour stages. Identifying scenarios in which a model's performance is less reliable may inform implementation. For example, the highly variable performance of the machine learning prediction models in **Chapter 5** was not detectable in the summary validation metrics. These sub-group analyses were appropriate but not exhaustive.

The expectation that a model will work uniformly well, or at least 'acceptably' (however that is defined) in all different societal groups may be unrealistic. A corollary is the concept of 'strong validation', defined as the assessment of calibration at every possible combination of predictor levels[24,25] – this is often computationally implausible. No model is ever the 'true model', and all modelling involves sampling and uncertainty. Predictor transformations, predictor selection or hyperparameter tuning, methods to handle missing

data, and model complexity or algorithmic approach can all contribute to variation in model performance at the individual level[26]. Further, predictor distributions and their associations within relevant sub-groups can vary, meaning that the aforementioned 'average effect' may not translate. When using small datasets to develop models, these factors can collectively manifest as instability on application to external datasets. When using larger datasets, this can manifest as variation in the predicted risks for a given individual from different approaches to modelling (see **Chapter 4**)[26,27]. Recent work has recommended a multi-level hierarchy of model stability to be assessed, namely the stability in a model's mean predicted risks, stability in a model's predicted risk distribution, stability in a model's predictions in sub-groups, and stability of model predictions for individuals[26]. This was not undertaken in this thesis due to the analyses being conducted prior to the methodology described in this work becoming available, but the bootstrap-based approaches would be considered in work beyond this thesis.

Whilst approaches exist to help the model handle such variation, such as including predictor interactions or stratification to permit different baseline risks in different settings, there should be a cognisance of sub-groups in which differences in model performance could be meaningful. Ideally, these should be pre-specified.

What is an appropriate validation?

Next, I consider what constitutes an appropriate validation. Whilst it is crucial to assess performance overall and in relevant groups, the applicability of results from validation to end implementation has been critically discussed recently. Some have stated that the term 'external validation' may be unhelpful, and instead the emphasis should be on

understanding and quantifying heterogeneity in performance, and monitoring performance over time and dynamically updating models as necessary[28]. Van Calster and colleagues invoke three arguments in support of this approach, namely: patient populations vary across different settings, measurements of predictors or outcomes vary, and the populations and measurements themselves vary over time[28]. The latter concept has been discussed in terms of 'clinician and dataset shift' by others[29], and the miscalibration of models described over varying timescales in other evaluation studies[30]. This thesis did not undertake an external evaluation of all developed models in an independent population – this was due to limited availability (and costs) of appropriate datasets. The UK Biobank analyses have limitations in terms of their applicability to the target population[16], as discussed in **Chapter 7**, and therefore the emphasis was on estimating the incremental effects of integrating additional data points into phenotypic models.

Within the internal-external validation framework[9] used in Chapters **3-5**, the types of variation outlined by van Calster and colleagues[28] were considered. Predicting with certainty the performance of a model developed using currently available or retrospective data on application to future populations is not possible, but simulating the development and application of models across geographically and temporally different data could provide an indication of how susceptible a model may be to degradation in performance in the near future.

For the models developed in **Chapters 3-5 & 7,** outcome rates have changed in recent decades, will likely continue to do so in the future, and new treatments will emerge. All of these will affect baseline risks. Together with an ageing, increasingly comorbid population in which risk factor distributions are evolving (e.g. reduced smoking or

increasing rates of obesity over time), changes in model performance may not always be predictable, but should be actively anticipated.

The results for the breast cancer mortality (**Chapters 5 & 7**) and breast cancer prognosis models (**Chapter 4**) may not apply to all future settings if they are ever implemented. External validation does not provide a rubber stamp of universal applicability[28]. Methods to dynamically update a prediction model should be considered as part of an implementation plan should any model be deployed in the real world[28] – these include temporal recalibration (updating the baseline survival function), landmarking, or periodic refitting of the model (i.e. updating all model coefficients) [31].

Does a model lead to better outcomes?

This thesis has focused on developing clinical prediction models and assessing their performance in terms of several metrics. Techniques such as decision curve analysis consider the effects of using a model on possible clinical decision making through a function of balancing true positives and false positives across a meaningful range of probabilities at which decisions might be made. Whilst it can provide support for a model having clinical usefulness (or the converse, as seen for some machine learning models in **Chapters 3-5**), it doesn't quantify the full effect of integrating a model into a public health system or clinical workflow, or what the best model-informed strategies are. Put another way, a population level stratification model may have excellent discrimination, good calibration and stable performance, but what decisions should be made using the model outputs and when?

Model-informed medical decision making, or model-informed decisions around screening for example, may have downstream effects. Further, if implementing the model incurs costs, it is imperative to not only quantify the benefits and harms from model-informed strategies, but to understand if these benefits are cost-effective.

To move from an impression of statistical performance towards an understanding of the effects of model implementation, randomised impact assessments can be used. Other techniques could also include decision analysis modelling[32] – this health economic evaluation approach is able to synthesise information from multiple sources (e.g. cohort studies and trials) regarding costs, benefits and utilities within a mathematical framework, and aims to provide decision makers (e.g. commissioners or regulatory bodies) with the best possible evidence for decision making whilst accounting for uncertainty. The approaches to do so range in complexity from decision trees to Markov cohort models, to individual-level microsimulations and discrete event simulations, and key metrics include, but are not limited to, incremental cost-effectiveness ratios and net monetary benefit.

The clinical impact and cost-effectiveness of risk-based screening has been simulated in several papers[33-35], but as summarised in **Chapter 1**, these tend to assess relatively simplistic stratification mechanisms or strategies, such as using single risk factors, rather than more nuanced multivariable prediction model approaches to estimating individual risk. Similar studies for example, have evaluated the clinical impact and cost-effectiveness of uniform multi-gene testing for all women diagnosed with breast cancer[36].

Taking the competing risks model from **Chapter 5**, for example, a key step in building the evidence base to support or reject an implementation within a population screening program could be a decision analysis model. Using individual level data from the population-representative data used to derive and evaluate the model (QResearch) and

combining this with published information regarding mammography and treatment costs, distributions of staging associated with different intensities of screening, and utilities associated with different health states, a health economic model could be used to simulate the effects of different risk-based strategies[32]. This could then also be extended to assess an important limitation of this thesis. **Chapter 7** reported that including a genome-wide polygenic risk score could provide moderate improvements in the prognostic ability of the competing risks model from **Chapter 5** but could not quantify whether this difference would manifest as improved clinical outcomes in a way that is cost-effective. This needs to consider the likely added costs of genomic sequencing all eligible women for a screening program. This is an active area of work underway seeking to build on this thesis.

Trusting a model – transparency, explainability, perceptions and fairness

Even if a model is robustly evaluated for implementation in its target population, has support for its use from decision analysis modelling, and is approved as a medical device by the relevant regulatory body, model adoption may still be limited in target end users. One of these factors is the problem of transparency.

The current state of play in clinical prediction modelling leverages associations between variables and an outcome – *"can this collection of information be useful to make accurate predictions?"* This is typically not within an explanatory framework[37-40] and variables can be a mixture of causal or non-causal factors or chance associations. As such, a model's coefficients cannot be causally interpreted[37-41] and in the case of 'black-box' methods or large ensemble models, can be difficult to visualise or understand at all[42]. Misinterpretation of coefficients in regression models as causal effects for example is not

uncommon[41]. Whilst methods to develop and evaluate models that function within causal or counterfactual frameworks (e.g. accounting for the effect of treatments in those that would be 'flagged' by an algorithm) are of interest, this is a nascent field undergoing further methodological development[39,43-45]. Predictors can be selected purely for their strength of association, selection distorts inference, and predictive models optimise for fitting the data best rather than explaining the true 'effect' of any given variable. The issue of transparency is therefore a multi-faceted one – understanding how a model generates its predictions is essential for a provider to make the decision and then relay the reasoning to the recipient, but human interpretations of the mechanics of a model can be inappropriate and lead to incorrect conclusions. The TRIPOD guidelines recommend that a clinical prediction model be presented in full for transparency[46] (e.g. for further study such as validation in other settings), and it can be useful to understand the different weights or contributions of certain factors to a risk score. But there is a difference between transparency and explainability – transparency refers to a user understanding how the predicted risk has been generated (e.g. though visualisation of coefficients, or techniques such as SHAP values for non-regression models), whereas explainability poses a more complex question regarding 'why' these variable values and combinations have led to the output[47,48]. For example, 'trust' in an algorithm can be eroded if a clinician thinks that the effect on the risk calculation of one variable is counterintuitive. Whether an artificial intelligence model can ever be explainable with current approaches, or if it needs to be, has been debated recently[42,49].

The reason that a parameter has its impact on the risk calculation may be any of the aforementioned types of 'effect'. Alternatively, and more concerning, it may reflect biases in the underlying data generating mechanism, such as under-representation, or structural inequalities[50,51] affecting access to care, diagnosis, or healthcare system

engagement. Indeed, there is significant concern regarding algorithmic bias[47,50-52] – properties in the model development data that percolate into the model and can then exacerbate or compound inequalities due to biased decision making. Bias in the model can lead to non-uniform utility, or even engender distrust in its outputs.

In **Chapter 5**, there was consideration of including ethnicity (a protected characteristic) as a predictor in a model to predict the risk that a woman would develop and then die from breast cancer within the next 10-years. To briefly recapitulate, when using 'White' as the reference category in regression modelling, the coefficients for most non-White groups were negative. These coefficients were not (and should not be) interpreted causally, but the perception of these model coefficients to end-users (e.g. women from ethnic minority backgrounds) raises a potential issue of explainability and trust. The 'reason' for this pattern of coefficients is complicated to disentangle, and end-users could perceive this as racist – e.g. even if all other characteristics were the same between a White woman and a woman from an Indian background, a White woman would be prioritised for screening or prevention. The work in that section of the thesis was transparent in reporting the model, but explaining the outputs to a user is complex. This pattern of ethnicity-related coefficients in regression models could be due to factors such as differences in ascertainment of the outcome, or the modelling of a 'combined' risk trajectory introducing complex associations with either incidence and/or death represented in a single model coefficient.

If there are differences between ascertainment of breast cancer diagnosis and/or death between ethnic groups, this could affect the direction and or magnitude of the association with the outcome modelled. Including ethnicity in the final model could embed and perpetuate this bias. For example, if women from a given ethnic group had the same true rate of breast cancer mortality as the others, but there was under-ascertainment of breast

cancer death in the development data, this could be 'captured' by the model in having a negative coefficient resulting in inappropriately low risk estimates, which would bias this group away from receiving interventions that reduce breast cancer mortality.

Using ethnicity as a predictor in a risk prediction model raises other issues. First, ethnic categories vary by country, which could affect model implementation. Second, race definitions can have logic that is rooted in little more than helpfulness for colonial administration – the diversity of groups within the 'Indian' Office for National Statistics category is vast, so what purpose would a single coefficient for this group serve? Third, ethnicity is a rough proxy for multiple factors ranging from genetic variation to structural racism, so including it in a model requires careful consideration. Including ethnicity as a predictor in COVID-19 prediction models partially formalised the increased risk of death observed across ethnic minority groups in a way to prioritise non-pharmacological interventions and vaccination decisions[53], which could be argued is a positive step, but the reason for ethnic minority groups having increased risks was not their skin colour.

Whilst there may be reasons to recognise variation in risk across certain groups (e.g. higher mortality from prostate cancer in Black men), one should be cautious[54]. Poorer outcomes may be due to historic discrimination rather than a true effect of ethnicity[52,54,55], including protected characteristics may lead to deselection from clinical strategies based on them, and their categorisation is not always helpful or logical. Future work should consider the roles, scope for and appropriateness of including ethnicity in prediction models. Standard prediction model coefficients cannot be interpreted causally, but perceptions of how a model seems to work or ways in which it may be biased can undermine trust and implementation[47].

Conclusions

This thesis explored the use of different algorithmic approaches to develop clinical prediction models that could support better decision making in breast cancer screening, prevention and management. The models that pertain to risk-based screening and prevention were further assessed in terms of the effects of integrating a genome-wide polygenic risk score. This was found to have some improvements in discrimination but how this translates into clinical impact and whether strategies informed by such models are cost-effective are to be elucidated. Further study after this thesis could focus on estimating the clinical impact and cost-effectiveness of strategies for screening/prevention based around the competing risks model from **Chapter 5**, and breast cancer management strategies that could be informed by the final models from **Chapter 4**. This, in conjunction with the growing body of evidence from qualitative work to understand acceptability of risk-based/model-informed strategies in these areas, and the effects of receiving risk estimates on programme participation[56-65] may inform the best approaches to later clinical implementation.

Chapter references

1. van Royen FS, Moons KGM, Geersing GJ, et al. Developing, validating, updating and judging the impact of prognostic models for respiratory diseases. *Eur Respir J* 2022; **60**(3).

2. Hueting TA, van Maaren MC, Hendriks MP, et al. The majority of 922 prediction models supporting breast cancer decision-making are at high risk of bias. *J Clin Epidemiol* 2022; **152**: 238-47.

3. Zheng Y, Li J, Wu Z, et al. Risk prediction models for breast cancer: a systematic review. *BMJ Open* 2022: 12:e055398.

4. Andaur Navarro CL, Damen JA, Takada T, et al. Systematic review finds "Spin" practices and poor reporting standards in studies on machine learning-based prediction models. *J Clin Epidemiol* 2023.

5. Dhiman P, Ma J, Andaur Navarro CL, et al. Overinterpretation of findings in machine learning prediction model studies in oncology: a systematic review. *J Clin Epidemiol* 2023.

6. Dhiman P, Ma J, Andaur Navarro CL, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. *BMC Med Res Methodol* 2022; **22**(1): 101.

7. Dhiman P, Ma J, Navarro CA, et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. *J Clin Epidemiol* 2021; **138**: 60-72.

8. Hudda MT, Archer L, van Smeden M, et al. Minimal reporting improvement after peer review in reports of COVID-19 prediction models: systematic review. *J Clin Epidemiol* 2023; **154**: 75-84.

9. Austin PC, van Klaveren D, Vergouwe Y, et al. Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. *J Clin Epidemiol* 2016; **79**: 76-85.

10. Louro J, Posso M, Hilton Boon M, et al. A systematic review and quality assessment of individualised breast cancer risk prediction models. *Br J Cancer* 2019; **121**(1): 76-85.

11. 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**(3): e007825.

12. 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**(8): e0160966.

13. Kerlikowske K, Chen S, Golmakani MK, et al. Cumulative Advanced Breast Cancer Risk Prediction Model Developed in a Screening Mammography Population. *J Natl Cancer Inst* 2022; **114**(5): 676-85.

14. Collins R. What makes UK Biobank special? *Lancet* 2012; **379**(9822): 1173-4.

15. Sperrin M, Riley RD, Collins GS, et al. Targeted validation: validating clinical prediction models in their intended population and setting. *Diagn Progn Res* 2022; **6**(1): 24.

16. Swanson JM. The UK Biobank and selection bias. *Lancet* 2012; **380**(9837): 110.

17. 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**(1): 21-34.

18. Evans DGR, van Veen EM, Harkness EF, et al. Breast cancer risk stratification in women of screening age: Incremental effects of adding mammographic density, polygenic risk, and a gene panel. *Genet Med* 2022; **24**(7): 1485-94.

19. 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**(4): 476-82.

20. Behravan H, Hartikainen JM, Tengstrom M, et al. Predicting breast cancer risk using interacting genetic and demographic factors and machine learning. *Sci Rep* 2020; **10**(1): 11044.

21. Thompson DJ, Wells D, Selzam S, et al. UK Biobank release and systematic evaluation of optimised polygenic risk scores for 53 diseases and quantitative traits. *MedRxiv* 2022.

22. Austin PC, van Klaveren D, Vergouwe Y, et al. Validation of prediction models: examining temporal and geographic stability of baseline risk and estimated covariate effects. *Diagn Progn Res* 2017; **1**: 12.

23. van Staa TP, Gulliford M, Ng ES, et al. Prediction of cardiovascular risk using Framingham, ASSIGN and QRISK2: how well do they predict individual rather than population risk? *PLoS One* 2014; **9**(10): e106455.

24. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019; **17**(1): 230.

25. Van Calster B, Nieboer D, Vergouwe Y, et al. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol* 2016; **74**: 167-76.

26. Riley RD, Collins GS. Stability of clinical prediction models developed using statistical or machine learning methods. *ArXiv* 2022.

27. Li Y, Sperrin M, Belmonte M, et al. Do population-level risk prediction models that use routinely collected health data reliably predict individual risks? *Sci Rep* 2019; **9**(1): 11222.

28. Van Calster B, Steyerberg EW, Wynants L, et al. There is no such thing as a validated prediction model. *BMC Med* 2023; **21**(1): 70.

29. Finlayson SG, Subbaswamy A, Singh K, et al. The Clinician and Dataset Shift in Artificial Intelligence. *N Engl J Med* 2021; **385**(3): 283-6.

30. Simpson CR, Robertson C, Kerr S, et al. External validation of the QCovid risk prediction algorithm for risk of COVID-19 hospitalisation and mortality in adults: national validation cohort study in Scotland. *Thorax* 2022; **77**(5): 497-504.

31. Booth S, Riley RD, Ensor J, et al. Temporal recalibration for improving prognostic model development and risk predictions in settings where survival is improving over time. *Int J Epidemiol* 2020; **49**(4): 1316-25.

32. Petrou S, Gray A. Economic evaluation using decision analytical modelling: design, conduct, analysis, and reporting. *BMJ* 2011; **342**: d1766.

33. Khan SA, Hernandez-Villafuerte KV, Muchadeyi MT, et al. Cost-effectiveness of risk-based breast cancer screening: A systematic review. *Int J Cancer* 2021.

34. Arnold M. Simulation modeling for stratified breast cancer screening - a systematic review of cost and quality of life assumptions. *BMC Health Serv Res* 2017; **17**(1): 802.

35. Arnold M, Pfeifer K, Quante AS. Is risk-stratified breast cancer screening economically efficient in Germany? *PLoS One* 2019; **14**(5): e0217213.

36. Sun L, Brentnall A, Patel S, et al. A Cost-effectiveness Analysis of Multigene Testing for All Patients With Breast Cancer. *JAMA Oncol* 2019; **5**(12): 1718-30.

37. Ramspek CL, Steyerberg EW, Riley RD, et al. Prediction or causality? A scoping review of their conflation within current observational research. *Eur J Epidemiol* 2021; **36**(9): 889-98.

38. van Diepen M, Ramspek CL, Jager KJ, et al. Prediction versus aetiology: common pitfalls and how to avoid them. *Nephrol Dial Transplant* 2017; **32**(suppl_2): ii1-ii5.

39. van Geloven N, Swanson SA, Ramspek CL, et al. Prediction meets causal inference: the role of treatment in clinical prediction models. *Eur J Epidemiol* 2020; **35**(7): 619-30.

40. Schooling CM, Jones HE. Clarifying questions about "risk factors": predictors versus explanation. *Emerg Themes Epidemiol* 2018; **15**: 10.

41. Westreich D, Greenland S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *Am J Epidemiol* 2013; **177**(4): 292-8.

42. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health* 2021; **3**(11): e745-e50.

43. Lin L, Sperrin M, Jenkins DA, et al. A scoping review of causal methods enabling predictions under hypothetical interventions. *Diagn Progn Res* 2021; **5**(1): 3.

44. Sperrin M, Diaz-Ordaz K, Pajouheshnia R. Invited Commentary: Treatment Drop-in-Making the Case for Causal Prediction. *Am J Epidemiol* 2021; **190**(10): 2015-8.

45. Sperrin M, Martin GP, Sisk R, et al. Missing data should be handled differently for prediction than for description or causal explanation. *J Clin Epidemiol* 2020; **125**: 183-7.

46. Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). *Ann Intern Med* 2015; **162**(10): 735-6.

47. Kundu S. AI in medicine must be explainable. *Nat Med* 2021; **27**(8): 1328.

48. Jalali MS, DiGennaro C, Sridhar D. Transparency assessment of COVID-19 models. *Lancet Glob Health* 2020; **8**(12): e1459-e60.

49. Reddy S. Explainability and artificial intelligence in medicine. *Lancet Digit Health* 2022; **4**(4): e214-e5.

50. Ganapathi S, Palmer J, Alderman JE, et al. Tackling bias in AI health datasets through the STANDING Together initiative. *Nat Med* 2022; **28**(11): 2232-3.

51. Seyyed-Kalantari L, Zhang H, McDermott MBA, et al. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med* 2021; **27**(12): 2176-82.

52. Vickers A. Do not treat Bill Gates for prostate cancer! Algorithmic bias and causality in medical prediction. *BJU Int* 2023; **131**(3): 263-4.

53. Clift AK, Coupland CAC, Keogh RH, et al. Living risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study. *BMJ* 2020; **371**: m3731.

54. Vickers AJ, Elfiky A, Freeman VL, Roach M, 3rd. Race, Biology, Disparities, and Prostate Cancer. *Eur Urol* 2022; **81**(5): 463-5.

55. Vickers AJ, Mahal B, Ogunwobi OO. Racism Does Not Cause Prostate Cancer, It Causes Prostate Cancer Death. *J Clin Oncol* 2023: JCO2202203.

56. French DP, Astley S, Brentnall AR, et al. What are the benefits and harms of risk stratified screening as part of the NHS breast screening Programme? Study protocol

for a multi-site non-randomised comparison of BC-predict versus usual screening (NCT04359420). *BMC Cancer* 2020; **20**(1): 570.

57. French DP, McWilliams L, Bowers S, et al. Psychological impact of risk-stratified screening as part of the NHS Breast Screening Programme: multi-site non-randomised comparison of BC-Predict versus usual screening (NCT04359420). *Br J Cancer* 2023; **128**(8): 1548-58.

58. French DP, McWilliams L, Howell A, et al. Does receiving high or low breast cancer risk estimates produce a reduction in subsequent breast cancer screening attendance? Cohort study. *Breast* 2022; **64**: 47-9.

59. French DP, Woof VG, Ruane H, et al. The feasibility of implementing risk stratification into a national breast cancer screening programme: a focus group study investigating the perspectives of healthcare personnel responsible for delivery. *BMC Womens Health* 2022; **22**(1): 142.

60. Gareth Evans D, McWilliams L, Astley S, et al. Quantifying the effects of risk-stratified breast cancer screening when delivered in real time as routine practice versus usual screening: the BC-Predict non-randomised controlled study (NCT04359420). *Br J Cancer* 2023: 1-9.

61. McWilliams L, Evans DG, Payne K, et al. Implementing Risk-Stratified Breast Screening in England: An Agenda Setting Meeting. *Cancers (Basel)* 2022; **14**(19).

62. McWilliams L, Woof VG, Donnelly LS, et al. Risk stratified breast cancer screening: UK healthcare policy decision-making stakeholders' views on a low-risk breast screening pathway. *BMC Cancer* 2020; **20**(1): 680.

63. Usher-Smith JA, Hindmarch S, French DP, et al. Proactive breast cancer risk assessment in primary care: a review based on the principles of screening. *Br J Cancer* 2023: 1-11.

64. Woof VG, Ruane H, French DP, et al. The introduction of risk stratified screening into the NHS breast screening Programme: views from British-Pakistani women. *BMC Cancer* 2020; **20**(1): 452.

65. Woof VG, Ruane H, Ulph F, et al. Engagement barriers and service inequities in the NHS Breast Screening Programme: Views from British-Pakistani women. *J Med Screen* 2020; **27**(3): 130-7.

Appendix 1 – Clinical code groups used to define model predictor variables (searchable on the QWeb webpage).

SNOMED = systematized nomenclature of medicine clinical terms; OPCS = Operating procedure codes supplement; ICD = international classification of diseases.

Variable	Read/SNOMED code lists used	Drug code groups	OPCS code lists used	ICD-10/ICD-9 code lists used
Smoking status	_smoking109_129 _nonsmoke_129 _lightsmoke_129 _modsmoke_129 _heavysmoke_129 _ex_smoker_129			
Polycystic ovary syndrome	_pcos_129			_pcos_icd10_129 _pcos_icd9
Hypertension	_htn_129			_hypertension_icd10_129 _hypertension_icd9
Ischaemic heart disease	_ischaemichd_129			_ihd_icd10_129 _ihd_icd9
Lung cancer	_lungca_129			_lungca_icd10_129 _lungca_icd9
Gynaecological cancers: Ovarian cancer Uterine cancer Endometrial cancer	 _ovarianca_129 _uterineca_129 _endometrialca_129			 _ovarianca_icd10_129 _ovarianca_icd9 _uterineca_icd10_129 _uterineca_icd9 _endometrialca_icd10_129
Breast cancer	_breastca_ox129			_breastca_icd10_129
Haematological cancer	_bloodca_129			_bloodca_icd10_129 _bloodca_icd9

Thyroid cancer	_thyroidca_129	_thyroidca_icd10_129 _thyroidca_icd9
Cirrhosis/chronic liver disease	_cirrhosis_cld_129	_cirrhosis_icd10_129 _cld_icd9
Type 1 diabetes mellitus	_t1dm_129	_t1dm_icd10_129
Type 2 diabetes mellitus	_t2dm_129	_t2dm_icd10_129
Chronic kidney disease Stage 3 Stage 4 Stage 5/ESRF/transplant/dialysis	 _ckd_3_129 _ckd_4_129 _ckd_5_129 _dialysistransplant_129	
Family history of gynaecological cancer: Family history of ovarian cancer Family history of uterine cancer Family history of female genital tract cancer	 _fh_ovarian_129 _fh_uterine_129 _fh_gentract_129	
Alcohol intake	_alco_intake_129	
Body mass index	_bmindex_129	
Systemic lupus erythematosus	_lupus_129	_lupus_icd10_129 _lupus_icd9
Vasculitis	_vasculitis_129	_vasculitis_icd10_129 _vasculitis_icd9
Endometriosis	_endometriosis1_129	_endometriosis_icd10_129 _endometriosis_icd9
Psychosis	_psychoses_129	_psychosis_icd10_129 _psychosis_icd9
Fibrocystic and benign breast disease	_fibro_breast_129	_fibrocystic_icd10_129 _benignbreast_icd9
Menopause	_menopause_129	
Breast carcinoma in situ	_breastcis_ox129	_breast_insitu_icd10_129 _breast_insitu_icd9
Family history of breast cancer	_fh_breastca_129	_fh_breast_icd10_129

Thaizides	_thiazide
Beta-blockers	_betablocker
Renin angiotensin axis antagonists	_reninheader
Angiotensin converting enzyme inhibitors	_ace
Calcium channel blockers	_cablocker
Antipsychotic drugs	_antipsychotic _antipsychotic_depot
Tricyclic and related antidepressant drugs	_tca
Monoamine oxidase inhibitors	_maoi
Other antidepressant drugs	_otherantidepressant
Selective serotonin reuptake inhibitor	_ssri
Combined oral contraceptive pill	_cop
Hormone replacement therapy	_hrt

Radiotherapy	_radiotherapy11
Mastectomy	_mastectomy11
Chemotherapy	_chemo_ox129
Hysterectomy	_hysterect_129
Other breast surgery	_breastsurgerynotmast_ox129

Appendix 2 – Sample size calculation code

Regarding the study by Paredes-Aracil, et al. "A scoring system to predict breast cancer mortality at 5 and 10 years" (https://www.nature.com/articles/s41598- 017-00536-7)

This study included 287 women diagnosed with breast cancer. The outcomes modelled were 5- and 10-year breast cancer mortality.

Using data in the publication, the following assumptions were made:

Breast cancer mortality rate 222/10,000 person years = annual rate 0.0222

Number of predictor parameters in final model = 9

Mean follow-up = 8.6 years

Prediction horizons = 5 and 10 (years)

15% of maximum permitted Cox-Snell R^2 permitted in this scenario = 0.09560669

Minimum sample size for both models = 802 (EPP=17.01) versus 287 used in study

R code (uses pmsampsize packaged from CRAN):

```
test5 <- pmsampsize(type="s", rsquared = 0.09560669, parameters = 9, timepoint = 5, meanfup = 8.6, rate=0.0222)

test10 <- pmsampsize(type="s", rsquared = 0.09560669, parameters = 9, timepoint = 10, meanfup = 8.6, rate=0.0222)
```