Supplementary Material of "Development and External Validation of Prediction Models for 10-Year Survival of Invasive Breast Cancer. Comparison with PREDICT and CancerMath."

Solon Karapanagiotis*, Paul D. P. Pharoah†, Christopher H. Jackson*, Paul J. Newcombe*

1 Methods

1.1 Description of model development

Markov chain Monte Carlo (MCMC) has become a very important computational tool in Bayesian statistics since it allows for Monte Carlo approximation of complex posterior distributions where analytical or numerical integration techniques are not applicable. Regarding variable selection Green, 1995 (24) demonstrated how classical MCMC methodology can be extended to explore models of differing dimensions using a 'Reversible Jump' algorithm (RJMCMC). Newcombe et al, 2017 (19) implemented a RJMCMC in a survival analysis context. Bayesian variable selection is well established to provide more stable selections of covariates compared to simpler stepwise methods, particularly when there are numerous and correlated variables to search over (19,25). Under this framework posterior inference is made on the predictors, and subsets of predictors, most likely associated with outcome. Attractive features include inference of probabilities for each predictor, and posterior inference on the model space. We present two implementations of the algorithm. The first using a Weibull model for the baseline survival time, and the second with a logistic regression model for the binary outcome of 10-year survival.

We start by noting that baseline variables age, whether the patient was detected via a screening programme, chemotherapy treatment, hormone therapy, the number of positive lymph nodes and tumour size were excluded from the model selection framework and fixed to be included in the model at all times. All baseline variables where transformed as in the latest version of PREDICT (7). Moreover, the effects of chemotherapy and hormone therapy were constrained to the effects reported for standard anthracycline-based chemotherapy and adjuvant tamoxifen from an updated analysis of the Early Breast Cancer Trialists Collaborative Group (22, 26), as was done previously in the development of the PREDICT model (2,7). Otherwise, the estimates of these treatment effects from the observational data would be affected by selection bias. Continuous variables were standardized and categorical were mean-centred to improve mixing of the MCMC chains.

*MRC Biostatistics Unit, University of Cambridge, Cambridge, UK
†Department of Oncology, University of Cambridge, Cambridge, UK

Subsequently, all pairwise interactions between variables were created and the RJMCMC method searched over those interactions.

1.1.1 The Weibull model

As noted above, we utilise both a Logistic and a Weibull model. Here we go in more depth into the Weibull regression model. Under the Weibull model, a patient i's hazard at time t is modelled as dependent on P covariate values, denoted by vector x_i, as

$$\lambda_i(t) = te^{\alpha + x_i\boldsymbol{\beta}},$$

where $\boldsymbol{\beta}$ is a P-length vector of covariate effects, and α denotes an intercept term. The survival function with a parameter k is:

$$S(t) = e^{-(t\lambda_i)^k},$$

$k > 0$, known as the Weibull 'shape' parameter. Let vector δ denote the log-hazard ratios associated with the 'fixed effects' i.e., the log-hazard ratios associated with the baseline covariates, and vector z_i denote the corresponding covariate values for patient i. Going forward, vector x_i will be used to denote patient i's interaction terms only, and vector β the log hazard ratios. P, the length of $\boldsymbol{\beta}$ therefore now denotes the number of interaction terms we wish to perform variable selection over.

Under Reversible Jump, variable selection is facilitated by placing a prior density on $\boldsymbol{\beta}$ which depends on a latent binary vector $\boldsymbol{\gamma} = (\gamma_1 \dots \gamma_p)$ of indices indicating whether each interaction is included in the model. For covariate p, $\gamma_p = 1$ indicates inclusion in the model. Conditional on the latent variable $\boldsymbol{\gamma}$, i.e. that a specific selection of interactions are included in the model, patient i's hazard may now be written as

$$\lambda_i(t)|\boldsymbol{\gamma} = e^{\alpha + \boldsymbol{z}_i\boldsymbol{\delta} + \boldsymbol{x}_{i,\gamma}\boldsymbol{\beta}_\gamma},$$

where vector $\boldsymbol{\beta}_\gamma$ contains only the non-zero elements of $\boldsymbol{\beta}$, and $\boldsymbol{x}_{i,\gamma}$, contains patient i's corresponding subset of interaction terms. The non-zero coefficients are assigned independent normal priors centred on 0, with a common variance σ_β^2,

$$p(\beta_p|\gamma_p = 1, \sigma_\beta^2) = N(0, \sigma_\beta^2)$$

for $p = 1 \dots P$. We chose a relatively informative Uniform(0.05, 2) prior for σ_β. This supports values for σ_β of maximum of 2, which corresponds to a prior with a 95% credible interval supporting hazard ratios between 0.02 and 50.9 – well outside the range we realistically expect to observe.

The 'fixed effects' $\boldsymbol{\delta}$ were ascribed weakly informative $N(0, 10^6)$ priors.

As mentioned, the effects of chemotherapy and hormone therapy were fixed. This was achieved using degenerative priors

$$p(\delta_{chemo}) = N(-0.249, 10^{-6})$$

$$p(\delta_{hormone}) = N(-0.416, 10^{-6}).$$

The model selection framework is completed by choosing a prior for $\boldsymbol{\gamma}$, the selection of the interactions included in the model. We used a beta-binomial prior

$$p(\boldsymbol{\gamma}) = \int p(\boldsymbol{\gamma}|\omega)p(\omega)d\omega = \frac{B(p_\gamma + a_\omega, P - p_\gamma + b_\omega)}{B(a_\omega, b_\omega)},$$

where B is the beta function and p_γ is the number of non-zero elements in $\boldsymbol{\gamma}$. Conceptually, ω denotes the underlying probability that each covariate has a non-zero effect, i.e. is included in $\boldsymbol{\gamma}$. Conditional on ω, all models of the same dimension are assumed, under this setup, equally likely a priori. a_ω and b_ω parametrise a beta hyper-prior on ω. We set $a_\omega = 1$ and $b_\omega = 32$ for ER-positive and $b_\omega = 13$ for ER-negative tumours, respectively. The higher values of b_ω relative to a_ω are introduced to encourage sparsity. This results in a prior on the probability of a true effect centred at $\approx 3\%$ for ER-positive and $\approx 8\%$ for ER-negative tumours. Finally, we must specify priors for the intercept α and the Weibull shape parameter k

$$p(\alpha) = N(0, 10^6)$$

$$p(log(k)) = N(0, 10^6).$$

A sensitivity analysis on σ_β, a_ω, b_ω, α, and k showed no substantive difference in estimates and inference.

1.1.2 The Logistic model

Let Y be the binary response variable with $Y_i = 0$ if the individual is alive at 10 years. Using the same notation as in the previous section the probability patient i is not alive at 10 years ($Pr(Y_i = 1)$) is

$$Pr(Y_i = 1|\boldsymbol{\gamma}) = \frac{e^{\alpha + \boldsymbol{z}_i\boldsymbol{\delta} + \boldsymbol{x}_{i,\gamma}\boldsymbol{\beta}_\gamma}}{1 + e^{\alpha + \boldsymbol{z}_i\boldsymbol{\delta} + \boldsymbol{x}_{i,\gamma}\boldsymbol{\beta}_\gamma}},$$

Now, the vectors β and δ correspond to log odds ratios. All priors remain as described previously expect the degenerative priors for the chemotherapy and hormone therapy effects which need to

be converted to odds ratios, i.e.,

$$p(\delta_{chemo}) = N(-0.353, 10^{-6})$$

$$p(\delta_{hormone}) = N(-0.534, 10^{-6}).$$

1.1.3 Model fitting

The Reversible Jump MCMC algorithm is an iterative sampling scheme. It starts at an initial model and corresponding set of parameter values, chosen at random. A new model and parameter values are proposed in each iteration. These are accepted with a probability proportional to their likelihood (logistic or weibull) and priors. The sampling stops when convergence to the target posterior distribution is achieved. Convergence was assessed using autocorrelation plots of the variable selections and chain plots of parameter values over the RJMCMC iterations.

Table S1: Fractional polynomial functions for age at diagnosis (years), tumour size (mm), number of positive nodes, tumour grade (I, II, III) and mode of detection (screening vs clinical) by ER status (positive, negative).

Prognostic factor	Function	
	ER positive	ER negative
Age1	$= \frac{[(age/10)^{-2} - 0.0287] - 0.004}{0.016}$	$= \frac{[age - 56.3254] + 0.0180}{13.4896}$
Age2	$= \frac{[(age/10)^{-2} \times ln(age/10) - 0.0510] - 0.0034}{0.017}$	-
Size	$= \frac{[ln(size/100) + 1.5452] + 0.1710}{0.5980}$	$= \frac{[(size/100)^{1/2} - 0.5090] + 0.0204}{0.1432}$
Nodes	$= \frac{[ln((nodes+1)/10) + 1.3876] + 0.3939}{0.7602}$	$= \frac{[1/[(nodes+1)/10]^{1/2} - 1.72] + 0.5204}{0.8967}$
Grade	$= grade - 2.0210$	$= grade - 2.7112$
Detection	$= detection - 0.3027$	-

Detailed results, in addition to the fixed effects, interaction terms and key model parameters, are presented in Tables S2 and S3 for ER-positive tumours and Tables S4 and S5 for ER-negative tumours.

Table S2: Results, for the fixed effects and interaction terms from RJMCMC Logistic model for ER-positive tumours.

	PostProb[a]	Median OR[b]	CrI_Lower[c]	CrI_Upper[c]	Median OR_Present[d]	CrI_Lower_Present[d]	CrI_Upper_Present[d]	BF[e]
Intercept (α)		-2.46	-2.71	-2.31				
Age1		0.31	-1.92	1.17				
Age2		-0.22	-1.06	1.72				
Size		0.48	0.36	0.60				
Nodes		0.64	0.53	0.73				
Grade		0.81	0.65	0.98				
Detection		-0.23	-0.49	0.02				
Chemo		-0.35	-0.35	-0.35				
Hormone		-0.53	-0.53	-0.53				
Age1_INT_Age1	0.25	1.00	1.00	1.13	1.09	1.02	1.16	10.56
Age1_INT_Age2	0.05	1.00	1.00	1.07	1.07	0.95	1.13	1.59
Age1_INT_Size	0.02	1.00	1.00	1.00	0.96	0.87	1.04	0.55
Age1_INT_Nodes	0.01	1.00	1.00	1.00	1.01	0.92	1.08	0.40
Age1_INT_Grade	0.03	1.00	1.00	1.00	0.94	0.83	1.06	0.83
Age1_INT_Detection	0.04	1.00	1.00	1.00	0.96	0.73	1.20	1.26
Age1_INT_Chemo	0.02	1.00	1.00	1.00	0.97	0.83	1.13	0.77
Age1_INT_Hormone	0.03	1.00	1.00	1.00	1.06	0.90	1.27	0.87
Age2_INT_Age2	0.21	1.00	0.93	1.26	1.18	0.89	1.32	8.27
Age2_INT_Size	0.02	1.00	1.00	1.00	0.98	0.91	1.07	0.56
Age2_INT_Nodes	0.01	1.00	1.00	1.00	1.01	0.93	1.08	0.49
Age2_INT_Grade	0.02	1.00	1.00	1.00	0.95	0.84	1.07	0.79
Age2_INT_Detection	0.03	1.00	1.00	1.00	0.98	0.65	1.29	1.11
Age2_INT_Chemo	0.03	1.00	1.00	1.00	0.98	0.83	1.18	1.00
Age2_INT_Hormone	0.02	1.00	1.00	1.00	1.00	0.82	1.18	0.80
Size_INT_Size	0.01	1.00	1.00	1.00	0.98	0.93	1.02	0.31
Size_INT_Nodes	0.02	1.00	1.00	1.00	0.95	0.87	1.04	0.69
Size_INT_Grade	0.03	1.00	1.00	1.00	0.94	0.82	1.06	1.03
Size_INT_Detection	0.03	1.00	1.00	1.02	1.07	0.89	1.31	1.13
Size_INT_Chemo	0.03	1.00	1.00	1.00	0.99	0.85	1.19	0.90
Size_INT_Hormone	0.36	1.00	1.00	1.64	1.32	1.05	1.72	18.17
Nodes_INT_Nodes	0.02	1.00	1.00	1.00	1.05	0.98	1.12	0.77
Nodes_INT_Grade	0.03	1.00	0.99	1.00	0.94	0.77	1.04	1.04
Nodes_INT_Detection	0.16	1.00	1.00	1.35	1.21	1.02	1.48	6.20
Nodes_INT_Chemo	0.20	1.00	1.00	1.31	1.19	1.01	1.40	8.08
Nodes_INT_Hormone	0.06	1.00	1.00	1.14	1.12	0.96	1.32	1.89
Grade_INT_Detection	0.05	1.00	1.00	1.11	1.12	0.84	1.43	1.65
Grade_INT_Chemo	0.03	1.00	1.00	1.00	0.95	0.76	1.18	0.89
Grade_INT_Hormone	0.04	1.00	1.00	1.05	1.10	0.88	1.42	1.32
Detection_INT_Chemo	0.05	1.00	1.00	1.07	1.07	0.85	1.49	1.62
Detection_INT_Hormone	0.04	1.00	1.00	1.00	0.98	0.74	1.40	1.37
Chemo_INT_Hormone	0.06	1.00	1.00	1.17	1.15	0.89	1.43	1.88

[a] Marginal Posterior Probability of Inclusion in the model – may be interpreted as the posterior probability an association exists with survival, adjusted for all other covariates in the model.

[b] Odds ratio.

[c] Upper and lower limit of 95% Credible Intervals (CrI).

[d] Odds ratio conditional on inclusion in the model.

[e] Bayes factor – summarizes the evidence in the data for the covariate included in the model. BF $>$ 1 indicates evidence in favour of including the covariate.

Table S3: Results, for the fixed effects and interaction terms from RJMCMC Weibull model for ER-positive tumours.

	PostProb[a]	Median HR[b]	CrI_Lower[c]	CrI_Upper[c]	Median HR_Present[d]	CrI_Lower_Present[d]	CrI_Upper_Present[d]	BF[e]
Weibull Scale[f]		0.37	0.30	0.44				
Intercept (α)		-1.90	-2.01	-1.79				
Age1		0.53	0.24	0.81				
Age2		-0.57	-0.88	-0.28				
Size		0.46	0.36	0.55				
Nodes		0.55	0.48	0.62				
Grade		0.75	0.61	0.88				
Detection		-0.31	-0.60	-0.09				
Chemo		-0.25	-0.25	-0.25				
Hormone		-0.42	-0.42	-0.42				
Age1_INT_Age1	0.00	1.00	1.00	1.00	1.00	0.99	1.01	0.10
Age1_INT_Age2	0.00	1.00	1.00	1.00	1.00	0.97	1.02	0.16
Age1_INT_Size	0.01	1.00	1.00	1.00	1.00	0.94	1.08	0.21
Age1_INT_Nodes	0.00	1.00	1.00	1.00	0.97	0.94	1.02	0.16
Age1_INT_Grade	0.01	1.00	1.00	1.00	0.96	0.88	1.04	0.42
Age1_INT_Detection	0.02	1.00	1.00	1.00	1.00	0.85	1.16	0.51
Age1_INT_Chemo	0.01	1.00	1.00	1.00	0.99	0.88	1.14	0.40
Age1_INT_Hormone	0.01	1.00	1.00	1.00	1.06	0.96	1.23	0.43
Age2_INT_Age2	0.01	1.00	1.00	1.00	1.00	0.95	1.04	0.24
Age2_INT_Size	0.01	1.00	1.00	1.00	0.99	0.92	1.04	0.25
Age2_INT_Nodes	0.01	1.00	1.00	1.00	0.99	0.93	1.04	0.17
Age2_INT_Grade	0.02	1.00	1.00	1.00	0.95	0.83	1.05	0.50
Age2_INT_Detection	0.03	1.00	1.00	1.00	1.08	0.76	1.63	0.92
Age2_INT_Chemo	0.02	1.00	1.00	1.00	1.02	0.89	1.21	0.49
Age2_INT_Hormone	0.02	1.00	1.00	1.00	1.02	0.91	1.21	0.55
Size_INT_Size	0.01	1.00	1.00	1.00	0.98	0.95	1.02	0.17
Size_INT_Nodes	0.02	1.00	1.00	1.00	0.96	0.92	1.00	0.51
Size_INT_Grade	0.02	1.00	1.00	1.00	0.97	0.86	1.07	0.49
Size_INT_Detection	0.02	1.00	1.00	1.00	1.00	0.68	1.24	0.60
Size_INT_Chemo	0.01	1.00	1.00	1.00	0.99	0.87	1.10	0.45
Size_INT_Hormone	0.13	1.00	1.00	1.30	1.20	1.03	1.47	4.92
Nodes_INT_Nodes	0.00	1.00	1.00	1.00	1.00	0.97	1.03	0.12
Nodes_INT_Grade	0.03	1.00	0.99	1.00	0.94	0.84	1.02	0.92
Nodes_INT_Detection	0.18	1.00	1.00	1.35	1.21	1.03	1.44	6.97
Nodes_INT_Chemo	0.02	1.00	1.00	1.00	1.07	0.99	1.17	0.66
Nodes_INT_Hormone	0.04	1.00	1.00	1.08	1.10	1.00	1.25	1.28
Grade_INT_Detection	0.02	1.00	1.00	1.00	1.06	0.90	1.36	0.77
Grade_INT_Chemo	0.02	1.00	1.00	1.00	0.96	0.83	1.18	0.56
Grade_INT_Hormone	0.02	1.00	1.00	1.00	1.07	0.90	1.42	0.63
Detection_INT_Chemo	0.04	1.00	1.00	1.06	1.13	0.89	1.67	1.19
Detection_INT_Hormone	0.01	1.00	1.00	1.00	1.01	0.80	1.24	0.49
Chemo_INT_Hormone	0.04	1.00	1.00	1.11	1.18	0.92	1.66	1.28

[a] Marginal Posterior Probability of Inclusion in the model – may be interpreted as the posterior probability an association exists with survival, adjusted for all other covariates in the model.

[b] Hazard ratio.

[c] Upper and lower limit of 95% Credible Intervals (CrI).

[d] Hazard ratio conditional on inclusion in the model.

[e] Bayes factor – summarizes the evidence in the data for the covariate included in the model. BF $>$ 1 indicates evidence in favour of including the covariate.

[f] Log(Weibull shape), i.e., $log(k)$ of Weibull model.

Table S4: Results, for the fixed effects and interaction terms from RJMCMC Logistic model for ER-negative tumours.

	PostProb[a]	Median OR[b]	CrI_Lower[c]	CrI_Upper[c]	Median OR_Present[d]	CrI_Lower_Present[d]	CrI_Upper_Present[d]	BF[e]
Intercept (α)		-0.82	-0.97	-0.67				
Age1		0.05	-0.09	0.19				
Size		0.37	0.22	0.53				
Nodes		0.70	0.55	0.85				
Grade		1.42	0.31	3.07				
Chemo		-0.35	-0.35	-0.35				
Age1_INT_Age1	0.03	1.00	1.00	1.01	1.06	0.96	1.20	0.42
Age1_INT_Size	0.02	1.00	1.00	1.00	0.98	0.83	1.07	0.32
Age1_INT_Nodes	0.02	1.00	1.00	1.00	0.98	0.87	1.14	0.28
Age1_INT_Grade	0.05	1.00	0.98	1.00	0.99	0.44	2.59	0.73
Age1_INT_Chemo	0.04	1.00	0.98	1.00	0.95	0.72	1.13	0.54
Size_INT_Size	0.02	1.00	1.00	1.00	0.98	0.90	1.07	0.24
Size_INT_Nodes	0.03	1.00	0.98	1.00	0.94	0.83	1.05	0.47
Size_INT_Grade	0.07	1.00	0.83	1.00	0.92	0.19	1.53	0.91
Size_INT_Chemo	0.03	1.00	1.00	1.00	1.00	0.82	1.26	0.47
Nodes_INT_Nodes	0.03	1.00	1.00	1.00	0.96	0.86	1.09	0.40
Nodes_INT_Grade	0.14	1.00	0.16	1.00	0.47	0.05	1.09	2.09
Nodes_INT_Chemo	0.05	1.00	0.94	1.00	0.93	0.62	1.10	0.66
Grade_INT_Chemo	0.06	1.00	0.94	1.01	0.98	0.36	3.17	0.83

[a] Marginal Posterior Probability of Inclusion in the model – may be interpreted as the posterior probability an association exists with survival, adjusted for all other covariates in the model.

[b] Odds ratio.

[c] Upper and lower limit of 95% Credible Intervals (CrI).

[d] Odds ratio conditional on inclusion in the model.

[e] Bayes factor – summarizes the evidence in the data for the covariate included in the model. BF $>$ 1 indicates evidence in favour of including the covariate.

Table S5: Results, for the fixed effects and interaction terms from RJMCMC Weibull model for ER-negative tumours.

	PostProb[a]	Median HR[b]	CrI_Lower[c]	CrI_Upper[c]	Median HR_Present[d]	CrI_Lower_Present[d]	CrI_Upper_Present[d]	BF[e]
Weibull Shape[f]		-0.05	-0.14	0.04				
Intercept (α)		-0.62	-0.76	-0.48				
Age1		0.13	0.03	0.23				
Size		0.32	0.21	0.44				
Nodes		0.59	0.48	0.70				
Grade		1.42	0.40	3.11				
Chemo		-0.25	-0.25	-0.25				
Age1_INT_Age1	0.07	1.00	1.00	1.08	1.07	0.98	1.14	0.92
Age1_INT_Size	0.02	1.00	1.00	1.00	0.97	0.89	1.05	0.27
Age1_INT_Nodes	0.02	1.00	1.00	1.00	0.99	0.90	1.05	0.28
Age1_INT_Grade	0.05	1.00	0.99	1.01	1.00	0.61	1.52	0.72
Age1_INT_Chemo	0.08	1.00	0.82	1.00	0.87	0.62	1.05	1.19
Size_INT_Size	0.02	1.00	1.00	1.00	0.98	0.94	1.02	0.25
Size_INT_Nodes	0.08	1.00	0.91	1.00	0.93	0.84	1.00	1.15
Size_INT_Grade	0.06	1.00	0.92	1.00	0.96	0.38	1.37	0.89
Size_INT_Chemo	0.04	1.00	0.96	1.00	0.93	0.77	1.09	0.52
Nodes_INT_Nodes	0.03	1.00	0.99	1.00	0.95	0.88	1.02	0.42
Nodes_INT_Grade	0.17	1.00	0.28	1.00	0.53	0.21	1.09	2.59
Nodes_INT_Chemo	0.05	1.00	0.94	1.00	0.93	0.78	1.09	0.62
Grade_INT_Chemo	0.05	1.00	0.96	1.00	0.96	0.64	1.37	0.73

[a] Marginal Posterior Probability of Inclusion in the model – may be interpreted as the posterior probability an association exists with survival, adjusted for all other covariates in the model.

[b] Hazard ratio.

[c] Upper and lower limit of 95% Credible Intervals (CrI).

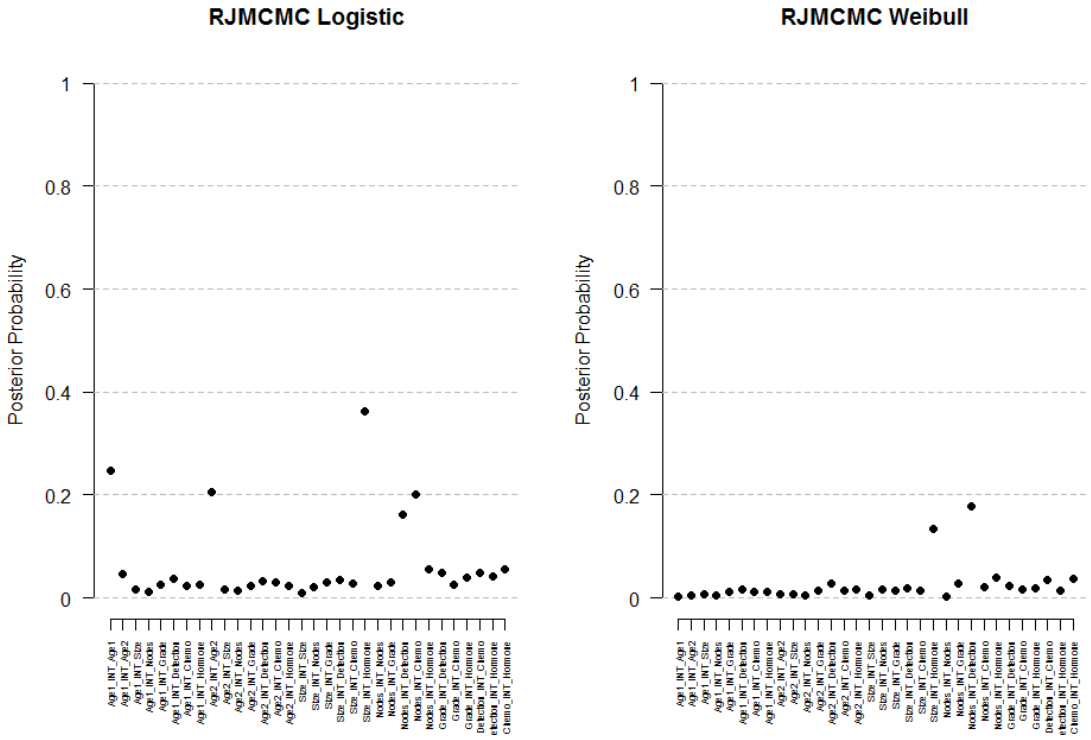
[d] Hazard ratio conditional on inclusion in the model.

[e] Bayes factor – summarizes the evidence in the data for the covariate included in the model. BF $>$ 1 indicates evidence in favour of including the covariate.

[f] Log(Weibull shape), i.e., $log(k)$ of Weibull model.

Figure S1: Marginal Posterior Probability of inclusion in the model for each interaction term, i.e. the relative frequency a covariate was included in the model for (A) ER positive and (B) ER negative tumours. It can be interpreted as the posterior probability an association exists with survival, adjusted for all other covariates in the model.

(A) ER positive



(B) ER negative

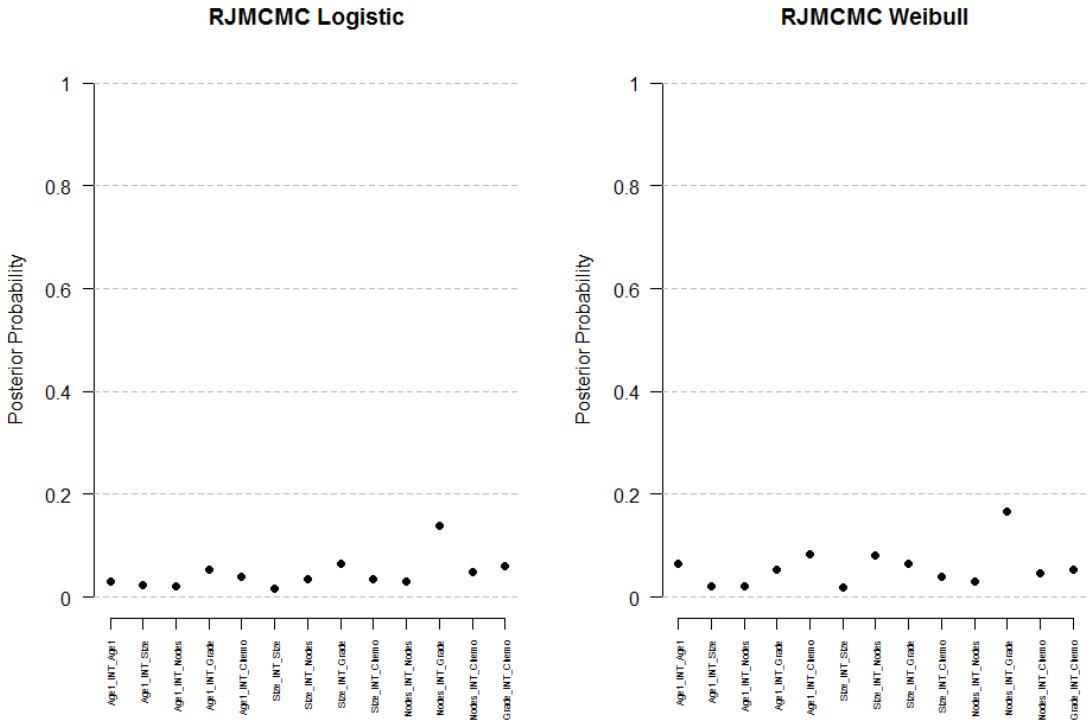
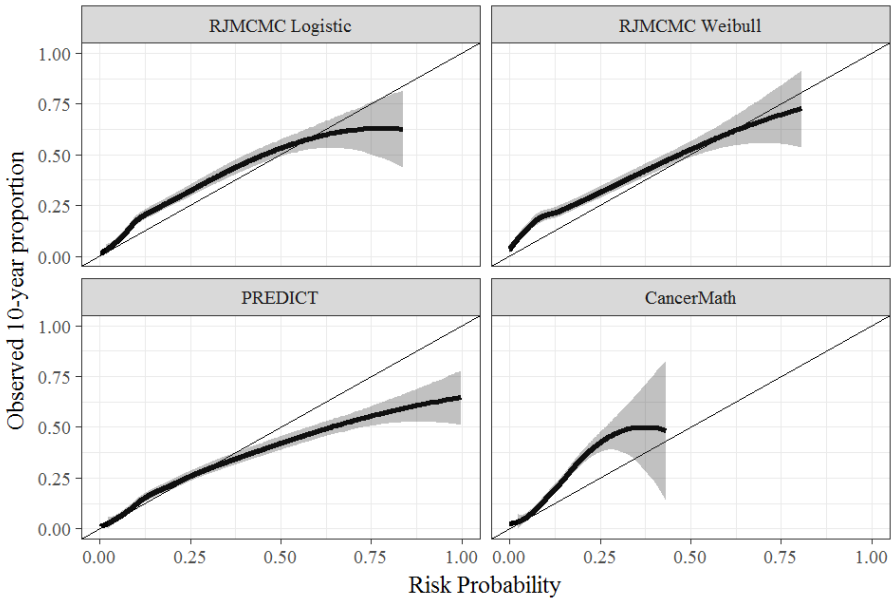


Figure S2: Individual calibration plots of the four models in the validation set with 95% CI for (A) ER positive and (B) ER negative tumours.

(A) ER positive



(B) ER negative

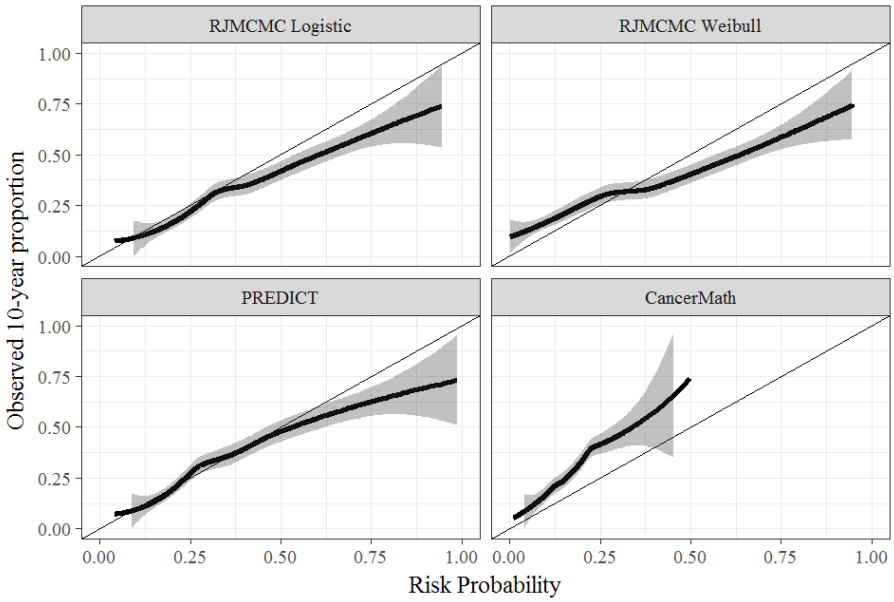
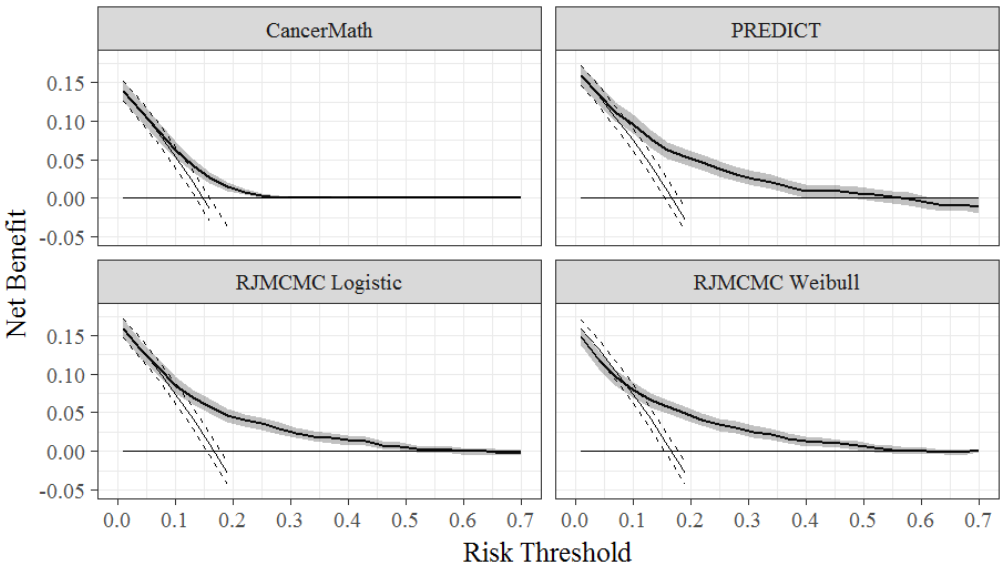


Figure S3: Individual net benefit plots of the four models in the validation set for (A) ER positive and (B) ER negative tumours. Shaded area represents 95% confidence intervals based on 1000 bootstrap replicates.

(A) ER positive



(B) ER negative