

Evaluating Temporal Dynamics in Breast Cancer Survival Predictions with Machine Learning and Cox Regression Analysis

Braden Woodhouse

University of Auckland <https://orcid.org/0000-0001-7162-6199>

Annette Lasham

University of Auckland <https://orcid.org/0000-0002-1084-4261>

Nicholas Knowlton

nknowlton@massey.ac.nz

Massey University <https://orcid.org/0000-0002-7022-3352>

Research Article

Keywords: Breast cancer, Survival prediction, Random Survival Forest, Cox proportional hazards, Machine learning, Tumour receptor subtype

Posted Date: March 4th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-5515692/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Accurate prediction of breast cancer-specific survival is crucial for guiding personalized treatment decisions and improving patient outcomes. This study evaluated the performance of machine learning approaches (Random Survival Forest, RSF and Generalized Boosted Model, GBM) alongside traditional Cox proportional hazards models for predicting survival in 21,574 women diagnosed with stage I-IV breast cancer in New Zealand between 2000-2019. Performance comparisons using time-dependent Area Under the Curve and Brier score metrics demonstrated that RSF consistently outperformed both Cox regression variants and GBM across all time points. Distinct differences emerged in survival predictions between modelling approaches: RSF captured a sharper initial decline in survival for most tumour receptor subtypes and better differentiated the favourable prognosis of ER+/HER2- tumours compared to other subtypes. Notably, variable importance analysis revealed fundamentally different prognostic emphases between modelling approaches—disease stage dominated Cox model predictions while tumour receptor subtype most strongly influenced RSF predictions. These findings highlight how machine learning approaches can capture complex, nonlinear relationships between clinical variables and survival outcomes that may be missed by traditional statistical models. The complementary insights provided by different modelling approaches suggest potential value in their combined use for enhanced risk stratification and more tailored treatment planning in breast cancer management, particularly when accounting for tumour biological characteristics alongside conventional staging factors.

Introduction

Breast cancer remains a global health priority, with one of the highest incidence and mortality rates among cancers affecting women (Amato 2023, Newman 2023). In New Zealand (NZ), breast cancer incidence is particularly high, with one in nine women receiving a diagnosis in their lifetime and significant disparities in outcomes persisting, especially among Māori and Pacific women (Te Aho o Te Kahu 2020, Kim 2025). While the average five-year survival rates in NZ have improved to 91%, there remains a pressing need to deepen our understanding of the factors influencing long-term survival and recurrence (Gautier 2022). Addressing this knowledge gap could drive more precise risk models and guide personalized follow-up and intervention strategies.

Traditionally, the Cox proportional hazards regression model has been instrumental in survival analysis, including in breast cancer research, by estimating hazard ratios (HRs) and highlighting risk factors (Lawrenson 2016, Elwood 2018). However, the Cox proportional hazards model relies on predefined predictor variables, which may constrain the discovery of complex, time-varying relationships. Machine learning (ML) approaches, particularly Random Survival Forests (RSF), offer an adaptable, data-driven alternative, by dynamically identifying important features and patterns within high-dimensional data. By allowing the data itself to determine critical prognostic factors, ML can identify interactions and nonlinear associations that may otherwise remain hidden (Cygu 2023, Mihaylov 2019).

Unlike traditional parametric models, ML approaches like RSF can flexibly handle nonlinear relationships, high-dimensional data, and variable interactions, making fewer assumptions about the data structure. For example, RSF techniques do not rely on a fixed baseline hazard, allowing for variable effects over time without requiring the proportional hazards assumption inherent to the Cox model (Xu 2022, Alafchi 2019). This flexibility is especially valuable when exploring nuanced relationships, such as those between receptor subtype and survival, that may not conform to proportional hazards.

In addition, ML methods like RSF incorporate automatic variable selection, reducing the manual effort required to select predictors. They often handle missing data more effectively than traditional methods, often employing ensemble techniques that enable predictions even when some data points are incomplete.

In this study, we chose not to employ extensions of the Cox proportional hazards model, such as those incorporating time-varying effects, due to the stringent assumptions required and their potential to limit applicability in long-term survival analysis (Bellera et al., 2010). Instead, we utilize RSF as an exemplar machine learning approach to explore survival and recurrence patterns among breast cancer patients in NZ. RSF serves as a complementary method, offering insights into complex prognostic factors while highlighting patterns not readily captured by traditional approaches. By leveraging both traditional statistical models and modern machine learning methods, this work provides a more comprehensive understanding of cancer outcomes and establishes a foundation for future advancements in personalized breast cancer care.

Results

Clinicopathological characteristics

Data for this study was sourced from Te Rēhita Mate Ūtaetae (Breast Cancer Foundation NZ National Register, subsequently referred to as Te Rēhita). Of 26,463 women diagnosed with stage 1–4 breast cancers in NZ between 2000 and 2019, 4,889 had missing information so were omitted (Fig. 1). This resulted in a study cohort of 21,574 women.

The majority of women were 45–69 years when their invasive breast cancer was diagnosed, were European, had ER+/HER2- tumour receptor subtype, had grade 2 tumours and early stage (1–2) disease (Table 1). Slightly more women had breast cancer detected after presenting with symptoms (approx. 55%) than through breast screening. Approximately half of these women underwent breast-conserving surgery (BCS), while the other half had a mastectomy. The majority (65%) also received adjuvant radiotherapy (Table 1). When analysed by tumour subtype, ER+/HER2- tumours were predominantly low-grade, whereas triple-negative tumours were predominantly high-grade. A higher proportion of women with ER+/HER2- breast tumours were over 44 years old and had their cancer diagnosed through screening (Table 1). These distinct differences in clinical and pathological features across receptor subtypes, which are already integral to clinical decision-making and treatment planning, highlight the

need for accurate, subtype-specific survival prediction models to further refine personalised clinical decision-making and treatment.

Table 1

Clinicopathological characteristics of the study cohort. A total of n = 21,574 women with invasive breast cancer and diagnosed between 2000–2019 were included in this study. Columns represent specific tumour receptor subtypes, which influence clinical decision making, and cohort characteristics are presented in rows.

Receptor Subtype					
	ER+/HER2-	ER+/HER2+	ER-/HER2+	Triple Negative	Overall
	(n = 16,177)	(n = 2,182)	(n = 1,087)	(n = 2,128)	(n = 21,574)
Age at Diagnosis (years)					
Mean (SD)	58.5 (12.3)	54.5 (13.1)	54.9 (13.0)	57.2 (14.1)	57.8 (12.7)
Median [Min, Max]	58.0 [20.0, 97.0]	53.0 [21.0, 94.0]	54.0 [22.0, 96.0]	57.0 [20.0, 98.0]	57.0 [20.0, 98.0]
Age at Diagnosis group (years)					
45–69	11,405 (70.5%)	1,396 (64.0%)	702 (64.6%)	1,276 (60.0%)	14,779 (68.5%)
≤ 44	1,863 (11.5%)	507 (23.2%)	237 (21.8%)	433 (20.3%)	3,040 (14.1%)
≥ 70	2,909 (18.0%)	279 (12.8%)	148 (13.6%)	419 (19.7%)	3,755 (17.4%)
Ethnicity					
European	12,026 (74.3%)	1,521 (69.7%)	714 (65.7%)	1,656 (77.8%)	15,917 (73.8%)
Māori	1,582 (9.8%)	237 (10.9%)	117 (10.8%)	158 (7.4%)	2,094 (9.7%)
Pacific Peoples	908 (5.6%)	177 (8.1%)	124 (11.4%)	82 (3.9%)	1,291 (6.0%)
Asian	1,390 (8.6%)	204 (9.3%)	104 (9.6%)	185 (8.7%)	1,883 (8.7%)
Other/Unknown	271 (1.7%)	43 (2.0%)	28 (2.6%)	47 (2.2%)	389 (1.8%)
Detection Method					
Screened	8,016 (49.6%)	795 (36.4%)	322 (29.6%)	594 (27.9%)	9,727 (45.1%)
Not Screened	8,161 (50.4%)	1,387 (63.6%)	765 (70.4%)	1,534 (72.1%)	11,847 (54.9%)

Receptor Subtype					
	ER+/HER2-	ER+/HER2+	ER-/HER2+	Triple Negative	Overall
Tumour Grade					
1	4,657 (28.8%)	106 (4.9%)	12 (1.1%)	35 (1.6%)	4,810 (22.3%)
2	8,621 (53.3%)	980 (44.9%)	233 (21.4%)	383 (18.0%)	10,217 (47.4%)
3	2,899 (17.9%)	1,096 (50.2%)	842 (77.5%)	1,710 (80.4%)	6,547 (30.3%)
Disease Stage (AJCC 7)					
1	9,868 (61.0%)	1,076 (49.3%)	466 (42.9%)	960 (45.1%)	12,370 (57.3%)
2	5,115 (31.6%)	827 (37.9%)	450 (41.4%)	944 (44.4%)	7,336 (34.0%)
3	877 (5.4%)	165 (7.6%)	89 (8.2%)	127 (6.0%)	1,258 (5.8%)
4	317 (2.0%)	114 (5.2%)	82 (7.5%)	97 (4.6%)	610 (2.8%)
Adjuvant Radiotherapy					
No	5,565 (34.4%)	763 (35.0%)	427 (39.3%)	749 (35.2%)	7,504 (34.8%)
Yes	10,612 (65.6%)	1,419 (65.0%)	660 (60.7%)	1,379 (64.8%)	14,070 (65.2%)
Surgery Type					
BCS	8,772 (54.2%)	853 (39.1%)	302 (27.8%)	980 (46.1%)	10,907 (50.6%)
Mastectomy	7,405 (45.8%)	1,329 (60.9%)	785 (72.2%)	1,148 (53.9%)	10,667 (49.4%)
Diagnosis Year Cluster					
2000–2004	853 (5.3%)	187 (8.6%)	139 (12.8%)	261 (12.3%)	1,440 (6.7%)
2005–2008	2,261 (14.0%)	298 (13.7%)	231 (21.3%)	414 (19.5%)	3,204 (14.9%)
2009–2019	13,063	1,697	717 (66.0%)	1,453	16,930

Receptor Subtype				
ER+/HER2-	ER+/HER2+	ER-/HER2+	Triple Negative	Overall
(80.8%)	(77.8%)		(68.3%)	(78.5%)

Machine learning model performance

To address this need, accurate survival models can increase our understanding of factors that affect breast cancer outcomes, inform treatment decision-making, and enable the stratification of patients into different risk groups, which is essential for personalised medicine. In order to generate an accurate and useful model, both machine learning and traditional models were explored for each receptor subtype. The survival prediction curves generated by the RSF model were contrasted with those generated by traditional adjusted CPH model, revealing notable differences in the shapes of the survival curves (Fig. 2). The RSF model predicted a sharper initial decline in survival across all except ER+/HER2- tumour receptor subtype, compared to the more gradual decline shown by the adjusted Kaplan-Meier (KM) curves. The RSF model predicted comparatively better survival for women with ER+/HER2- breast cancers relative to other receptor subtypes, whereas the curves for the adjusted KM are more condensed, potentially obscuring these distinctions.

Next, the time-dependent survival prediction performance was evaluated via AUC and Brier score for five different models (see methods). A high AUC and low Brier score is indicative of a better performing model, or closer survival prediction to the true values. This analysis showed that RSF was the optimal model for survival prediction across all time points (Fig. 3). The other models analysed had similar survival prediction performance.

Since the different Cox proportional hazards models studied had roughly equivalent prediction accuracy, for simplicity, this study proceeded with the simplest Cox model as the main comparison model representing a traditional tool. The Cox model with interactions and regularised Cox model were not explored further in this study. The traditional model CPH and RSF were then compared to evaluate the risk factors associated with breast cancer-specific survival (BCSS). To further understand the factors driving these predictions, the influential predictors identified through both models were analysed.

Influential predictors identified through traditional Cox regression and RSF models

The RSF model (which demonstrated the most accurate survival predictions) and the Cox proportional hazards model were analysed to uncover their underlying mechanisms and identify the key variables influencing their predictions. In the Cox proportional hazards model, worse BCSS was significantly associated with diagnosis from age 70, diagnosis after presenting with symptoms (i.e. not through

breast screening), having triple negative receptor subtype, higher tumour grade and disease stage, and requiring a mastectomy instead of BCS (Table 2). In contrast, Asian women, and those with the ER+/HER2 + receptor subtype were associated with improved BCSS compared to other ethnicities and tumour subtypes, respectively (Table 2).

Table 2
Cox proportional hazards model. Hazard ratios for breast cancer-specific survival are presented for each covariable, with 95% confidence intervals (CI) in parentheses.

		Hazard Ratio (95% CI)	P value
Age at diagnosis (years)	45–69	<i>Reference</i>	
	≤ 44	0.96 (0.86–1.07)	0.42
	≥ 70	1.41 (1.26–1.57)	0.00
Ethnicity	European	<i>Reference</i>	
	Māori	1.13 (0.99–1.29)	0.07
	Pacific Peoples	0.98 (0.83–1.14)	0.76
	Asian	0.6 (0.5–0.71)	0.00
	Other/Unknown	1.38 (1.08–1.75)	0.01
Detection method	Screened	<i>Reference</i>	
	Not Screened	1.52 (1.35–1.71)	0.00
Receptor Subtype	ER+/HER2-	<i>Reference</i>	
	ER+/HER2+	0.87 (0.76–0.99)	0.04
	ER-/HER2+	1.11 (0.96–1.3)	0.16
	Triple Negative	1.48 (1.31–1.66)	0.00
Tumour Grade	1	<i>Reference</i>	
	2	2.48 (2.05–3.01)	0.00
	3	4 (3.27–4.88)	0.00
Disease stage	1	<i>Reference</i>	
	2	2.26 (2.02–2.52)	0.00
	3	4.17 (3.57–4.87)	0.00
	4	11.33 (9.77–13.15)	0.00
Radiotherapy	No	<i>Reference</i>	
	Yes	0.97 (0.88–1.06)	0.47
Most invasive surgery	BCS	<i>Reference</i>	
	Mastectomy	1.45 (1.3–1.62)	0.00

		Hazard Ratio (95% CI)	P value
Diagnosis year cluster	2000–2004	<i>Reference</i>	
	2005–2008	0.72 (0.63–0.82)	0.00
	2009–2019	0.47 (0.42–0.53)	0.00

While the Cox proportional hazards model generates hazards ratios which can be used to assess the impact of variables of survival, RSF model does not, although it does provide a measure “variable importance”. RSF variable importance is different from statistical model coefficients, however, it provides a metric for comparison and an alternative tool when assessing which variables influence survival and survival prediction. This analysis showed that stage, grade, receptor subtype and surgery type were the four most influential risk factors in the RSF model, when considering permutation variable importance (Fig. 4A). These results were consistent with the statistically significant coefficients observed in the Cox proportional hazards model. Ethnicity was the next most important variable, which also demonstrated statistical significance in the Cox proportional hazards model.

Having identified the key predictors in each model, the next step was to explore how these variables contribute to the models' overall predictive performance. To achieve this, Brier score loss was compared after variable permutation. This analysis revealed that disease stage was the most important variable for the performance of the Cox proportional hazards model, whereas tumour receptor subtype was the most important variable for the performance of the RSF survival prediction model (Fig. 4B and C).

Discussion

This study demonstrates the utility of machine learning methods, particularly Random Survival Forests (RSF), in analysing BCSS alongside traditional statistical approaches such as Cox proportional hazards model. Our findings indicate that the RSF model provided the most accurate survival predictions across all time points, outperforming traditional models in this dataset. While RSF does not yield the familiar hazard ratios (HRs) associated with Cox proportional hazards model, it offers alternative metrics such as variable importance and Brier score loss after permutations to identify influential predictors of survival.

Analysis of the importance of the individual variables in the RSF model largely aligned with the significant HRs identified by the Cox model. Both models highlighted disease stage, tumour grade, receptor subtype, and surgery type as key predictors of BCSS. Notably, in the RSF model, receptor subtype emerged as the most influential predictor, contrasting with the Cox model where disease stage held the greatest influence. This difference highlights the potential of RSF to capture the complex, nonlinear relationships between receptor subtype and survival outcomes, which may be less apparent in traditional Cox regression analysis. This enhanced ability to detect intricate patterns underscores the potential of machine learning methods to uncover nuanced prognostic factors. This finding has

important implications for personalised medicine in breast cancer. While disease stage remains a critical factor in treatment decisions, the RSF model's emphasis on receptor subtype provides a complementary perspective. By accurately predicting survival based on subtype-specific factors, this model could lead to more tailored treatment decisions, particularly within specific disease stages. For instance, the model may identify specific subtypes with particularly favourable or unfavourable prognoses within a given stage, allowing for more informed treatment selection or closer surveillance. Overall, the RSF model's emphasis on receptor subtype, in contrast to the Cox model's focus on disease stage, highlights its potential to enhance prognostication and treatment decision-making in breast cancer, ultimately leading to improved patient outcomes.

An interesting observation from this study was that the performance of survival prediction for all models decreased as time from diagnosis increased, indicated by declining AUC and increasing Brier scores. Importantly, diagnosis year cluster was included in these models, which adjusts for some of the expected change over time, such as changes in treatment method. This trend may reflect model underfitting due to limited long-term data (11% of women in this cohort had follow-up times of 15 years or greater) or fewer events (deaths) occurring at extended follow-up times. However, performance metrics appeared to stabilize or even improve approaching the 20-year follow-up mark. This could suggest that the models are better at predicting long-term survivors, possibly due to distinct characteristics among women who survive beyond 20 years after diagnosis (1% of women in this cohort). These survivors may have unique clinical or biological features that are more readily captured by the models at extended time points. This pattern may also be influenced by survivorship bias, meaning the individuals who remain in the cohort at extended time points could represent a selective group with inherently better prognostic factors, potentially inflating model performance at those longer time points.

The strengths and limitations of both traditional and machine learning methods are evident in this study. Traditional methods like Cox proportional hazards model rely on strong assumptions about the data, such as the proportional hazards assumption, which can be violated in long-term survival data (Kurt Omurlu et al., 2009; Wang & Li, 2017). In contrast, machine learning models like RSF can model complex nonlinear relationships and interactions between covariates without stringent assumptions, providing flexibility in handling diverse data structures. This flexibility is demonstrated in our findings, where RSF consistently outperformed traditional models in prediction accuracy across all time points (Fig. 3). However, increased model complexity can reduce interpretability. In this study we began to explore the underlying mechanism of how the models generated their predictions (Fig. 4). Future studies could include techniques like Local Interpretable Model-Agnostic Explanations and Shapley Additive Explanation values that can explore the influence of variables in complex models even further (Alabi et al., 2023; Lundberg & Lee, 2017; Moncada-Torres et al., 2021).

Machine learning models also handle missing data more effectively and can produce predictions even with incomplete data through ensemble methods and the use of weak learners (Fanizzi et al., 2023; Steele et al., 2018). Women with missing data were excluded from this study (with the exception of ethnicity) to ensure comparability with the Cox proportional hazards model, which requires complete

datasets. Future studies could consider retaining these records to prevent information loss and explore statistical methods such as multiple imputation or other machine learning approaches for missing data.

Our findings align with previous studies comparing traditional and machine learning methods for survival analysis. Spooner et al. (2020) found similar performance across various machine learning algorithms and traditional Cox proportional hazards model, depending on whether model assumptions are met and the complexity of covariate relationships. In cases where the assumptions are violated or relationships are more complex, machine learning models may outperform traditional methods. Even when traditional model assumptions hold, the ability of machine learning models to detect more complex relationships provides an opportunity to enhance traditional models by identifying additional variables or interactions to include. Our study found better survival prediction performance for RSF compared to Cox proportional hazards, which aligns with the findings from Jia et al. (2025), although their study cohort included only inflammatory breast cancer patients. Studies exploring breast cancer survival or recurrence as a binary analysis (ignoring censoring) are common (Hamedi et al., 2024; Kamble et al., 2025; Noman et al., 2025). Noman et al. (2025) nicely utilised a Cox proportional hazards model alongside other machine learning models, however, this was to predict recurrence, whereas our study analysed BCSS, so we are unable to compare results.

It is important to note that there is no universally 'best' model across all datasets (Manikandan et al., 2023). The choice of model should be guided by the specific context, data characteristics, and research objectives. While machine learning offers advantages in modelling complex relationships, traditional models remain valuable, especially when their assumptions are appropriate for the data.

The utility and interpretation of survival prediction models should be approached cautiously, with rigorous validation and calibration studies, and in conjunction with clinical expertise. Individual variability among patients necessitates careful consideration when applying these models for personalized prognostication. However, both traditional statistical models and machine learning methods are valuable tools for exploring survival patterns and identifying influential predictors in patient populations.

Conclusion

This study underscores the potential of machine learning methods like RSF in enhancing survival analysis for breast cancer patients. The RSF model's ability to capture complex relationships without strict assumptions makes it a powerful complement to traditional methods. Key predictors such as tumour receptor subtype and disease stage were identified as influential for BCSS, highlighting the need for models that can accurately capture these complexities. Our findings advocate for the integration of machine learning approaches in survival analysis to improve risk stratification and support the development of personalised care strategies in breast cancer management.

Methods

Ethics

This study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Auckland Health Research Ethics Committee (AH2800). This study used data from Te Rēhita Mate Ūtaetae- the Breast Cancer Foundation NZ National Register. Te Rēhita is an opt-out register, which operates under the NZ Health and Disability Ethics Committee approval (16/NTA/139/AM03), privacy, and health legislation and Treaty of Waitangi principles. Patients receive an information sheet explaining that their de-identified data may be used for research purposes, subject to approval by Te Rēhita governance group. Those who choose not to opt-out implicitly consent to their data being included in this study.

Study cohort

This study analysed data from 26,463 women diagnosed with stage 1–4 breast cancers in NZ. A 2000-to-2019-time span was selected to ensure a follow-up period of at least 3 years, and to mitigate bias introduced by the inclusion of data from new regions in 2020. Patient follow-up within Te Rēhita was confirmed as up to date as of the data extraction on March 2, 2023. BCSS was calculated from the date of diagnosis to the date of death from breast cancer, if it occurred. Otherwise, survival time was censored at the date of death from other causes or at the latest follow-up date within the study period. Women in the study cohort had a median follow-up time of 8 years.

Predictors

This analysis incorporated several key patient factors as predictors, including age group at diagnosis, detection method, ethnicity, surgical intervention, and radiotherapy. Age groups were categorized as < 45, 45–69, and ≥ 70 years to correspond with the eligibility criteria for New Zealand’s national breast screening program, BreastScreen Aotearoa (BSA) (Breast Cancer Aotearoa Coalition, 2020) for women aged 45–69 years. Detection method was recorded as invasive breast cancer detected by screening mammography- “screened”, or detection after presentation with symptoms- “not screened”. Te Rēhita collects up to three ethnicities per person, and sources these from New Zealand’s Ministry of Health through an interactive link with each person’s unique health identifier (Gautier et al., 2022). Ethnicity was recorded as prioritised ethnicity, level 1 per HISO 10001:2017 ethnicity data protocols (Wellington: Ministry of Health, 2017), with Middle Eastern, Latin American, African (MELAA), “Other,” and missing ethnicity information combined into an “Other/Unknown” category. The most invasive surgery performed on each patient was recorded, ensuring that cases where a mastectomy followed an initial BCS were accurately captured. Disease stage was assessed using the AJCC 7 TNM staging system to ensure consistency in staging across the entire cohort. Year of diagnosis was grouped into intervals (2000–2004, 2005–2008, and 2009–2019) based on the results of k-medians survival clustering from our previous study (*reference clustering paper*).

Models

In this analysis, we evaluated several models for survival prediction, including standard Cox regression, regularized Cox regression, Cox regression with interactions, Random Survival Forests (RSF), and Generalized Boosting Models (GBM).

The Cox proportional hazards model (Cox, 1972) is a semi-parametric model (Kalbfleisch & Schaubel, 2023), containing components of both parametric (known distribution of regression coefficients β_i) and non-parametric (unknown baseline hazard function h_0). The Cox model takes the formula

$$h(t) = h_0(t) \bullet \exp(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)$$

where $h(t)$ is the hazard function at time t and h_0 is the baseline hazard with $x_i = 0$ for all p predictors. The hazard function is the conditional probability of a person experiencing an event at some time point, for example breast cancer-specific death, given that the person has been event-free up until that time. Hazard ratios can be calculated by taking the exponential of the coefficients $\exp(\beta_i)$, and are measures of an instantaneous relative risk (Sashegyi & Ferry, 2017). Note that the log of the hazard rate is a linear combination of covariates. The other assumption for the Cox model is that the effect of covariates on the hazard function is proportional over time, since the exponent $\exp(\beta_i x_i)$ does not include time t . That is, the difference in hazard for one group at one time point, maintains the same proportion difference in hazard to other groups at any other time point (Bewick et al., 2004; Kuitunen et al., 2021). In non-mathematical terms, the Cox model helps to quantify how each variable impacts the person's risk of experiencing the event.

Machine learning provides some additional useful tools to examine survival that do not share the same rigorous statistical assumptions as those just described for Cox regression. Random Survival Forests (RSF) (Ishwaran et al., 2008) are extensions of random forests (Breiman, 2001), ensemble tree methods that combine and average the survival predictions from many decision trees (Breiman et al., 1984). RSFs reduce estimation variance by using independent bootstrap sampling before constructing each tree, with each split chosen from a random subset of the features rather than all features. In the context of survival analysis, RSF extends this approach by using a splitting criterion optimized for survival differences, such as the log-rank test, to construct each tree (Ishwaran et al., 2008). Each node within an RSF tree divides the data to maximize survival contrast between groups, enabling the detection of complex, nonlinear relationships among covariates. Unlike Cox regression, RSF does not produce hazard ratios but instead provides aggregate survival predictions derived from many trees. To determine variable importance in RSF, we used an out-of-bag (OOB) approach, where the OOB data — data not included in the bootstrap sample — were permuted for each variable. Comparing prediction error before and after permutation indicated each variable's importance, with larger discrepancies in error signifying greater influence on survival. Variables appearing higher in the tree structure influence more downstream nodes, and therefore, permutations of high-importance variables typically result in larger prediction errors.

Generalised Boosted Models (GBM) works by consecutively building decision trees to predict the residuals of the previous tree (Ridgeway, 2020), similar to Gradient Boosting (Chen & Guestrin, 2016).

Model implementation

Cox regression was implemented in R using `survival::coxph()`. Regularized Cox regression used `riskRegression::GLMnet()`, with alpha values tuned over a grid from 0 to 1 in 0.1 increments, resulting in an optimal alpha of 0, which effectively applied ridge regularization. Lambda was selected based on the minimum prediction error using the default regularization path. By incorporating regularization, we anticipated improved model generalizability and enhanced survival prediction accuracy on unseen test data, as regularization mitigates overfitting. Additionally, a Cox model with interaction terms was developed to account for the interaction between disease stage and ethnicity, which was found to be significant during exploratory analysis.

RSF was tuned on all covariates from the multivariable model, along with the diagnosis year, using `randomForestSRC::tune.rfsrc`. The optimal configuration was achieved with `nodesize = 15` and `mtry = 6`. Feature importance scores and their confidence intervals were computed using delete-d jackknife procedures (`randomForestSRC::subsample`), while survival curves from RSF predictions were visualized through `ggRandomForests::gg_rfsrc()`.

The GBM model was fit in this study using `gbm::gbm()`, and employed 10-fold cross-validation over a parameter grid to minimize the AUC for 5-year BCSS. Optimal GBM parameters were identified as `n.tree = 300`, `interaction.depth = 3`, and `shrinkage = 0.1`, with covariate influence on survival assessed through relative variable importance scores.

To evaluate the survival prediction performance across models, we used the Area Under the Curve (AUC) and Brier score at 5-year follow-up and time-dependent intervals. The `riskRegression::Score()` function, which employs inverse probability of censoring weights (IPCW), was utilized for Brier score estimation. AUC, which quantifies a model's discriminatory power, reflects the probability that a randomly selected positive instance (breast-cancer-specific death) is ranked higher than a negative one (survival) (Fawcett, 2006). Time-dependent AUC was calculated using the Blanche et al. method, a modification of the Uno method (Uno et al., 2007), wherein each AUC point reflects the probability that a patient who died from breast cancer had a higher predicted risk than a patient who survived (Wu & Li, 2018). Confidence intervals for these metrics were generated through 10-fold cross-validation.

To enhance interpretability of the Cox regression and RSF models, we employed model-agnostic explanations using the `survex` package (Spytek et al., 2023), which allowed for time-dependent permutation-based feature importance. This approach provided insights into the dynamic influence of covariates over time and facilitated a deeper understanding of variable impacts on survival prediction in a model-agnostic context.

Declarations

Competing interests

The author(s) declare no competing interests.

Author Contributions

BW conducted the initial data analysis, contributed to study design, and participated in the interpretation of results. NK conceptualized and designed the study, interpreted the findings, and contributed to manuscript drafting. AL assisted in the interpretation of results and manuscript preparation. All authors reviewed and approved the final manuscript.

Acknowledgements

We gratefully acknowledge Te Rēhita Mate Ūtaetae - Breast Cancer Foundation National Register for providing the data used in this study and Te Rēhita Clinical Advisory Group for reviewing this manuscript. Our thanks also extend to all the individuals who consented to participate in Te Rēhita, contributing invaluable information towards breast cancer research and management. We would also like to thank our funders, the New Zealand Breast Cancer Foundation via the Helena McAlpine Young Women's Breast Cancer Study and the Breast Cancer Cure via the Not A One-Size-Fits-All Service grant.

Data availability

The dataset presented in this article is not readily available because this requires approval by the custodians of the data. Requests to access the dataset should be directed to Te Rēhita Mate Ūtaetae.

References

1. Amato, O., Guarneri, V., & Girardi, F. (2023). Epidemiology trends and progress in breast cancer survival: earlier diagnosis, new therapeutics. *Current Opinion in Oncology*, 35(6), 612.
<https://doi.org/10.1097/CCO.0000000000000991>
2. Arnold, M., Morgan, E., Rumgay, H., Mafra, A., Singh, D., Laversanne, M., Vignat, J., Gralow, J. R., Cardoso, F., Siesling, S., & Soerjomataram, I. (2022). Current and future burden of breast cancer: Global statistics for 2020 and 2040. *Breast (Edinburgh, Scotland)*, 66, 15–23.
<https://doi.org/10.1016/J.BREAST.2022.08.010>
3. Aye, P. S., Win, S. S., Tin Tin, S., & Elwood, J. M. (2023). Comparison of Cancer Mortality and Incidence Between New Zealand and Australia and Reflection on Differences in Cancer Care: An Ecological Cross-Sectional Study of 2014-2018. *Cancer Control: Journal of the Moffitt Cancer Center*, 30. <https://doi.org/10.1177/10732748231152330>
4. Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 12: Survival analysis. *Critical Care*, 8(5), 389.
<https://doi.org/10.1186/CC2955>
5. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
<https://doi.org/10.1023/A:1010933404324/METRICS>

6. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. In *Classification and Regression Trees*. CRC Press.
<https://doi.org/https://doi.org/10.1201/9781315139470>
7. Chih-Lin Chi, W Nick Street, & William H Wolberg. (2007). Application of artificial neural network-based survival analysis on two breast cancer datasets. *AMIA Annu Symp Proc* . <https://pubmed.ncbi.nlm.nih.gov.ezproxy.auckland.ac.nz/18693812/>
8. Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 187–202. <https://doi.org/10.1111/J.2517-6161.1972.TB00899.X>
9. Cygu, S., Seow, H., Dushoff, J., & Bolker, B. M. (2023). Comparing machine learning approaches to incorporate time-varying covariates in predicting cancer survival time. *Scientific Reports 2023 13:1*, 13(1), 1–10. <https://doi.org/10.1038/s41598-023-28393-7>
10. Elwood, J. M., Tawfiq, E., TinTin, S., Marshall, R. J., Phung, T. M., Campbell, I., Harvey, V., & Lawrenson, R. (2018). Development and validation of a new predictive model for breast cancer survival in New Zealand and comparison to the Nottingham prognostic index. *BMC Cancer*, 18(1), 897.
<https://doi.org/10.1186/s12885-018-4791-x>
11. Gautier, A., Harvey, V., Kleinsman, S., Knowlton, N., Lasham, A., & Ramsaroop, R. (2022). *30,000 voices: Informing a better future for breast cancer in Aotearoa New Zealand*. Breast Cancer Foundation NZ. <https://doi.org/10.17608/k6.auckland.19679019>
12. Gupta, S., Tran, T., Luo, W., Phung, D., Kennedy, R. L., Broad, A., Campbell, D., Kipp, D., Singh, M., Khasraw, M., Matheson, L., Ashley, D. M., & Venkatesh, S. (2014). Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. *BMJ Open*, 4(3), e004007. <https://doi.org/10.1136/BMJOPEN-2013-004007>
13. Hamedj, S. Z., Emami, H., Khayamzadeh, M., Rabiei, R., Aria, M., Akrami, M., & Zangouri, V. (2024). Application of machine learning in breast cancer survival prediction using a multimethod approach. *Scientific Reports 2024 14:1*, 14(1), 1–18. <https://doi.org/10.1038/s41598-024-81734-y>
14. Hao, J., Kim, Y., Mallavarapu, T., Oh, J. H., & Kang, M. (2019). Interpretable deep neural network for cancer survival analysis by integrating genomic and clinical data. *BMC Medical Genomics*, 12(Suppl 10). <https://doi.org/10.1186/S12920-019-0624-2>
15. Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *Https://Doi.Org/10.1214/08-AOAS169*, 2(3), 841–860. <https://doi.org/10.1214/08-AOAS169>
16. Jia, Y., Li, C., Feng, C., Sun, S., Cai, Y., Yao, P., Wei, X., Feng, Z., Liu, Y., Lv, W., Wu, H., Wu, F., Zhang, L., Zhang, S., & Ma, X. (2025). Prognostic prediction for inflammatory breast cancer patients using random survival forest modelling. *Translational Oncology*, 52, 102246.
<https://doi.org/10.1016/J.TRANON.2024.102246>
17. Kalbfleisch, J. D., & Schaubel, D. E. (2023). Fifty Years of the Cox Model. *Https://Doi.Org/10.1146/Annurev-Statistics-033021-014043*, 10, 1–23.
<https://doi.org/10.1146/ANNUREV-STATISTICS-033021-014043>

18. Kamble, T. S., Wang, H., Myers, N., Littlefield, N., Reid, L., McCarthy, C. S., Lee, Y. J., Liu, H., Pantanowitz, L., Amirian, S., Rashidi, H. H., & Tafti, A. P. (2025). Predicting cancer survival at different stages: Insights from fair and explainable machine learning approaches. *International Journal of Medical Informatics*, 197, 105822. <https://doi.org/10.1016/J.IJMEDINF.2025.105822>
19. Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282), 457–481. <https://doi.org/10.1080/01621459.1958.10501452>
20. Kim, J., Harper, A., McCormack, V. *et al.* Global patterns and trends in breast cancer incidence and mortality across 185 countries. *Nat Med* (2025). <https://doi.org/10.1038/s41591-025-03502-3>
21. Kuitunen, I., Ponkilainen, V. T., Uimonen, M. M., Eskelinen, A., & Reito, A. (2021). Testing the proportional hazards assumption in cox regression and dealing with possible non-proportionality in total joint arthroplasty research: methodological perspectives and review. *BMC Musculoskeletal Disorders*, 22(1), 1–7. <https://doi.org/10.1186/S12891-021-04379-2/TABLES/2>
22. Lawrenson, R., Lao, C., Elwood, M., Brown, C., Sarfati, D., & Campbell, I. (2016). Urban Rural Differences in Breast Cancer in New Zealand. *International Journal of Environmental Research and Public Health*, 13(10). <https://doi.org/10.3390/ijerph13101000>
23. Mihaylov, I., Nisheva, M., & Vassilev, D. (2019). Application of Machine Learning Models for Survival Prognosis in Breast Cancer Studies. *Information 2019, Vol. 10, Page 93*, 10(3), 93. <https://doi.org/10.3390/INFO10030093>
24. Montazeri, M., Montazeri, M., Montazeri, M., & Beigzadeh, A. (2016). Machine learning models in breast cancer survival prediction. *Technology and Health Care: Official Journal of the European Society for Engineering and Medicine*, 24(1), 31–42. <https://doi.org/10.3233/THC-151071>
25. Najafi-Vosough, R., Faradmal, J., Tapak, L., Alafchi, B., Najafi-Ghobadi, K., & Mohammadi, T. (2022). Prediction the survival of patients with breast cancer using random survival forests for competing risks. *Journal of Preventive Medicine and Hygiene*, 63(2), E298. <https://doi.org/10.15167/2421-4248/JPMH2022.63.2.2405>
26. New Zealand. Te Aho o te Kahu. (2020). *The state of cancer in New Zealand 2020*.
27. Newman, L. (2023). Oncologic anthropology: Global variations in breast cancer risk, biology, and outcome. *Journal of Surgical Oncology*, 128(6), 959–966. <https://doi.org/10.1002/JSO.27459>
28. Noman, S. M., Fadel, Y. M., Henedak, M. T., Attia, N. A., Essam, M., Elmaasarawii, S., Fouad, F. A., Eltasawi, E. G., & Al-Atabany, W. (2025). Leveraging survival analysis and machine learning for accurate prediction of breast cancer recurrence and metastasis. *Scientific Reports 2025 15:1*, 15(1), 1–16. <https://doi.org/10.1038/s41598-025-87622-3>
29. Pölsterl, S., Sarasua, I., Gutiérrez-Becker, B., & Wachinger, C. (2020). A wide and deep neural network for survival analysis from anatomical shape and tabular clinical data. *Communications in Computer and Information Science*, 1167 CCIS, 453–464. https://doi.org/10.1007/978-3-030-43823-4_37/COVER

30. Sashegyi, A., & Ferry, D. (2017). On the Interpretation of the Hazard Ratio and Communication of Survival Benefit. *The Oncologist*, 22(4), 484. <https://doi.org/10.1634/THEONCOLOGIST.2016-0198>

31. Wellington: Ministry of Health. (2017). HISO 10001:2017 Ethnicity Data Protocols. In *Ethnicity Data Protocols*. Ministry of Health. <chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.tewhatauora.govt.nz/assets/Our-health-system/Digital-health/Health-information-standards/HISO-10001-2017-Ethnicity-Data-Protocols.pdf>

Figures

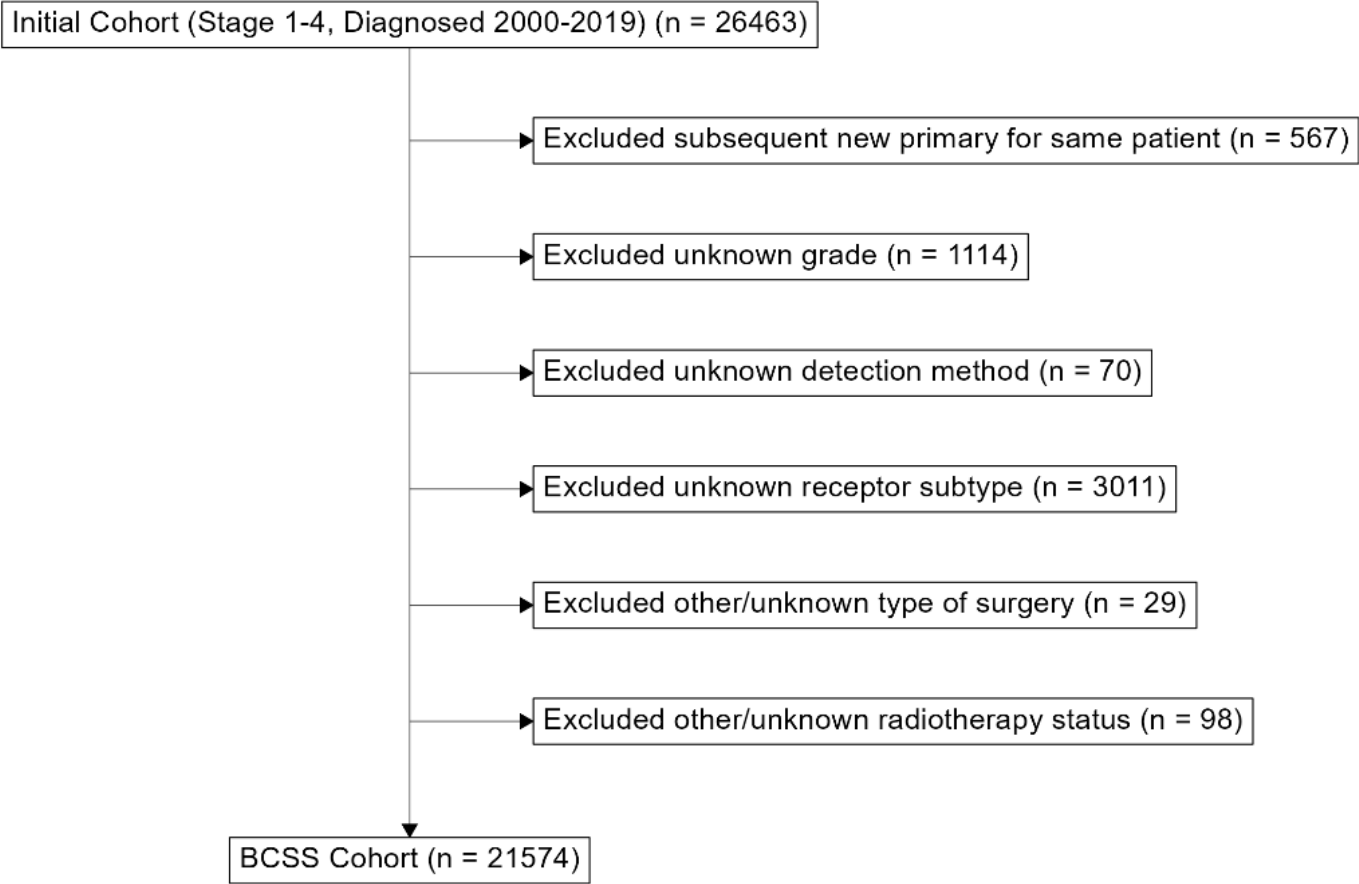


Figure 1

Consort diagram showing the number of women excluded in each category. This study analysed breast cancer-specific survival (BCSS) for women with invasive breast cancers (stage 1-4) diagnosed in the years 2000-2019 (n=21,574 following exclusions).

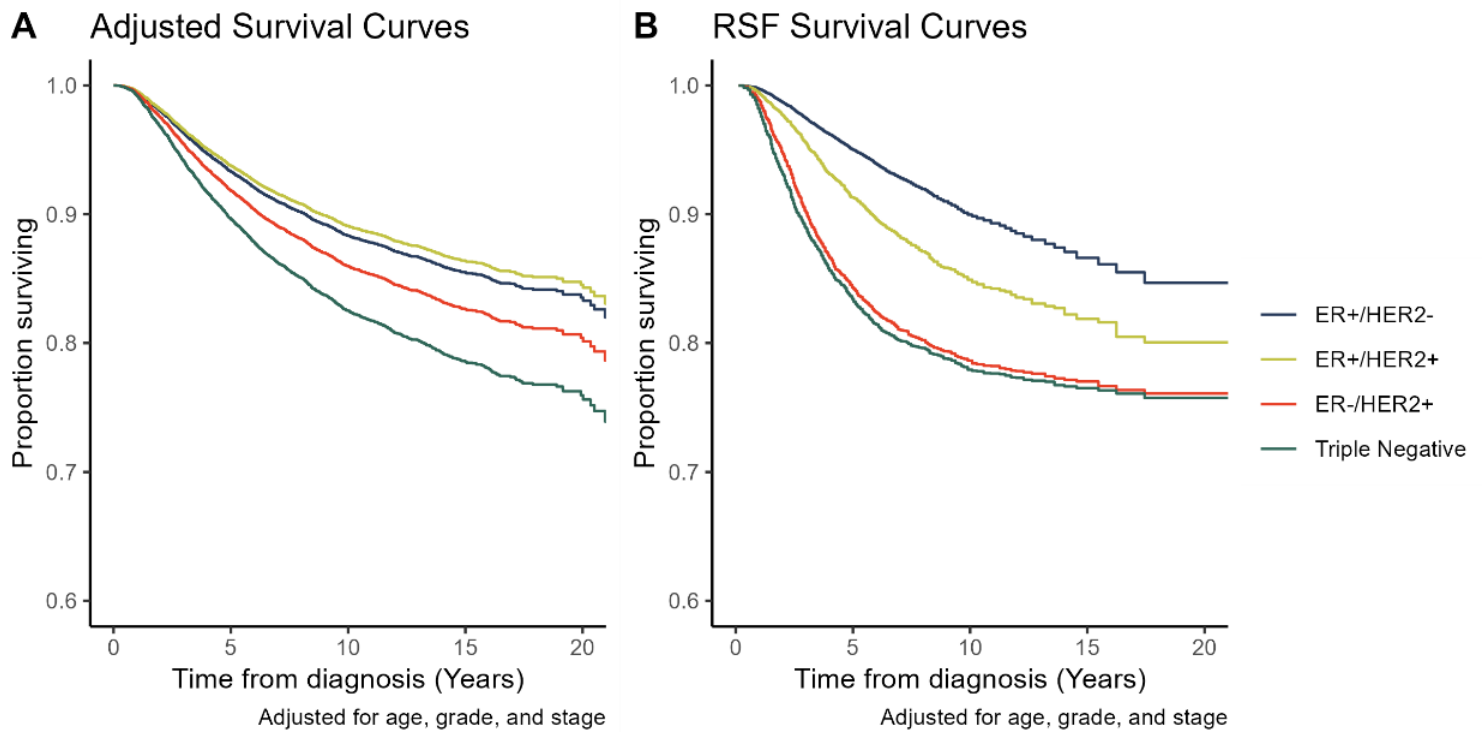


Figure 2

Proportion of women surviving with invasive breast cancers over time by receptor subtype. (A) Adjusted Kaplan-Meier curves using “direct” adjustment method and (B) predicted proportion of women surviving using random survival forest (RSF) model. All curves were adjusted for patient age, tumour grade and disease stage.

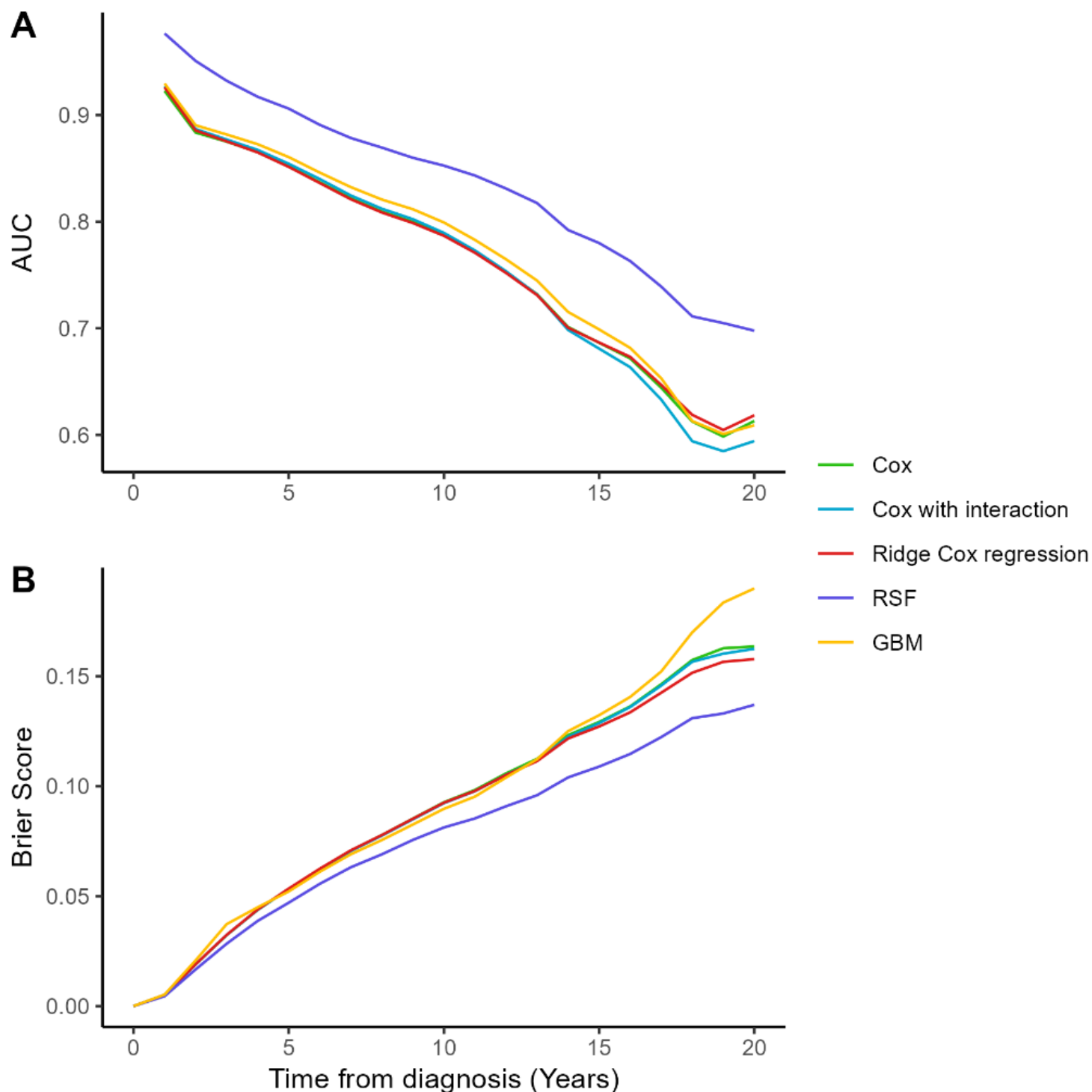


Figure 3

Model performance for prediction of breast cancer-specific survival. Cox proportional hazards model, Cox proportional hazards model with interaction terms between receptor subtype and grade and between diagnosis year cluster and detection method, regularised Cox proportional hazards model (ridge), Random Survival Forest (RSF) and Generalised Boosted Model (GBM) were evaluated for survival prediction performance over time using measures (A) area under the curve (AUC) and (B) Brier score.

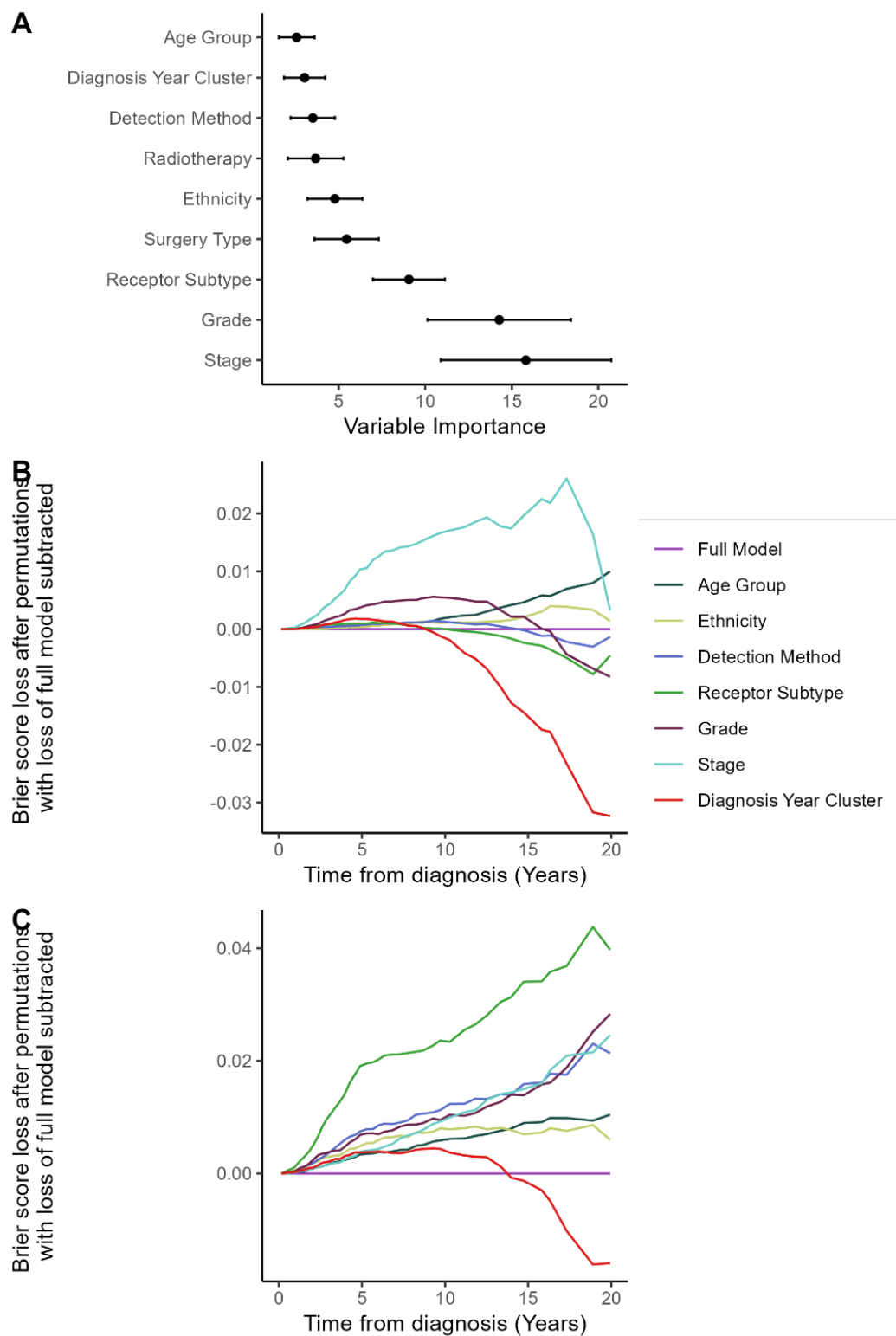


Figure 4

Analysis of variable importance. (A) Forest plot showing the most important variables associated with predicting BCSS in the RSF model. Error bars represent 95% confidence intervals generated using 100 jack-knife subsamples. Model-agnostic explanations for (B) Cox proportional hazards and (C) RSF models, showing permutation variable importance over time.