Technical report of the final exam of the course "Statistical models for Healthcare data"

Integrating Classification, Survival, and Causal Inference Models for Breast Cancer Prognosis

Anastasiia Samsonova

Abstract: Background: Breast cancer prognosis depends on a complex interplay of tumor characteristics, patient factors, and treatment strategies. Accurate risk stratification and reliable estimation of treatment effects are essential for clinical decision-making, yet remain challenging in observational settings due to non-randomized treatment assignment and confounding. **Methods:** Using data from the Rotterdam Breast Cancer Study, we applied complementary statistical modeling frameworks to investigate recurrence and all-cause mortality. Classification models based on regularized logistic regression were used to predict binary outcomes, while Cox proportional hazards models were employed to analyze recurrence-free and overall survival. In addition, a causal inference analysis was conducted to estimate the average treatment effect of adjuvant hormonal therapy using outcome regression, with adjustment for clinically relevant confounders. Model performance, assumptions, and interpretability were carefully assessed across analyses. **Results:** Across classification and survival analyses, tumor-related factors—particularly lymph node involvement, tumor grade, and tumor size—emerged as the most consistent prognostic indicators for both recurrence and mortality. Treatment-related variables, including hormonal therapy and chemotherapy, were associated with reduced risk, with stronger protective effects observed for recurrence-related outcomes than for all-cause mortality. Time-to-event models revealed time-dependent effects and residual departures from proportional hazards, reflecting the complexity of long-term survival modeling. Causal inference yielded more conservative estimates of the effect of hormonal therapy, with modest average risk reductions accompanied by substantial uncertainty. **Conclusions:** This study illustrates how predictive, survival, and causal inference approaches provide complementary insights into breast cancer prognosis and treatment effects. While predictive models and survival analyses offer robust tools for risk stratification, causal estimates highlight the limitations imposed by confounding and limited overlap in observational data. Together, these findings emphasize the importance of aligning statistical methodology with clearly defined analytical objectives in healthcare research.

1. Introduction

Breast cancer is a heterogeneous disease characterized by substantial variability in clinical course and long-term outcomes. Prognosis depends on a combination of patient characteristics, tumor biology, treatment strategies, and healthcare-related factors. Accurate risk stratification is therefore central to clinical decision-making, particularly in the context of adjuvant therapies aimed at reducing recurrence and mortality risk [1,2].

Statistical models play a key role in quantifying prognosis and supporting treatment decisions in oncology. Classification models are commonly used to predict binary outcomes such as recurrence or death, while survival models provide a dynamic perspective by explicitly accounting for time-to-event information and censoring [3]. However, estimating treatment effects in observational settings remains challenging due to non-randomized treatment assignment and the presence of confounding, particularly confounding by indication, which continues to represent a central methodological challenge in oncology research [4].

In this project, we analyze data from the Rotterdam Breast Cancer Study, a well-established cohort with extended follow-up, to investigate prognosis and treatment effects using complementary statistical modeling frameworks. The analysis integrates predictive modeling, survival analysis, and causal inference approaches, allowing for a comprehensive assessment of both prognostic factors and therapeutic effects within a single observational dataset.

The overall aim of this study is to evaluate prognosis and treatment effects in breast cancer patients using observational cohort data. Specifically, we pursue the following objectives:

- to identify key predictors of tumor recurrence and long-term mortality using classification models;
- to examine recurrence-free survival and overall survival through time-to-event analyses;
- to estimate the causal effect of adjuvant hormonal therapy on clinical outcomes while accounting for measured confounding.

By combining these analytical perspectives, this study highlights how different statistical frameworks provide complementary insights into breast cancer prognosis and treatment effects, while also underscoring the inherent limitations of causal interpretation in observational healthcare data.

2. Materials and Methods

2.1 Study Population and Outcomes

This study is based on data from the Rotterdam Breast Cancer Study, an observational cohort comprising 2982 patients diagnosed with primary breast cancer. Among these patients, 339 (11.4%) received adjuvant hormonal therapy, while 2643 (88.6%) did not receive hormonal treatment. The cohort includes patients treated across multiple hospitals, with extended follow-up allowing for the analysis of both short- and long-term clinical outcomes.

Baseline clinical and pathological variables collected at diagnosis include age at diagnosis, year of surgery, menopausal status, tumor size category, histological grade, number of positive lymph nodes, estrogen receptor (ER) and progesterone receptor (PGR) levels, as well as information on adjuvant treatments, including chemotherapy and hormonal therapy. These variables reflect key prognostic factors routinely used in breast cancer risk stratification and treatment decision-making. Baseline characteristics of the study population, stratified by hormonal therapy status, are summarized in Table 1; continuous variables are reported as median and interquartile range, while categorical variables are presented as percentages.

Table 1. Baseline characteristics of the study population stratified by hormonal therapy status.

Variable\Population	Overall	Hormonal therapy: Yes	Hormonal therapy: No
age (years)	54.0 [45.0, 65.0]	62.0 [57.0, 69.0]	53.0 [44.0, 64.0]
pgr (fmol/mg)	41.0 [4.0, 198.0]	19.0 [1.0, 115.0]	46.0 [5.0, 208.0]
nodes (number of nodes)	1.0 [0.0, 4.0]	4.0 [2.0, 9.0]	0.0 [0.0, 3.0]
er (fmol/mg)	61.0 [11.0, 202.8]	79.0 [14.0, 247.5]	59.0 [11.0, 197.0]
year (calendar year)	1988.0 [1986.0, 1990.0]	1991.0 [1989.0, 1992.0]	1988.0 [1986.0, 1990.0]
grade (category)	{'3': 73.4%, '2': 26.6%}	{'3': 82.6%, '2': 17.4%}	{'3': 72.2%, '2': 27.8%}
size (category)	{'<=20': 46.5%, '20-50': 43.3%, '>50': 10.2%}	{'20-50': 50.7%, '<=20': 30.7%, '>50': 18.6%}	{'<=20': 48.5%, '20-50': 42.3%, '>50': 9.1%}
meno (status)	{True: 56.0%, False: 44.0%}	{True: 87.9%, False: 12.1%}	{True: 51.9%, False: 48.1%}

Patients were followed from the time of surgery until the occurrence of the event of interest or censoring. Two types of outcomes were considered in this analysis. First, binary outcomes were defined to support classification modeling: tumor recurrence and all-cause mortality. Second, time-to-event outcomes were analyzed to account for censoring and variable follow-up times, including time to recurrence and time to death.

Specifically, tumor recurrence was defined as the first documented recurrence event during follow-up, with time measured in days from surgery to recurrence or censoring. Overall survival was defined as time from surgery to death from any cause or censoring. Patients who did not experience the event of interest during follow-up were treated as right-censored at their last observed time.

This combination of binary and time-to-event outcome definitions enables a complementary analytical strategy, allowing predictive performance to be assessed alongside time-dependent risk dynamics and treatment effects.

2.2 Covariates and Data Preprocessing

Covariates were selected based on clinical relevance, data availability, and methodological considerations. Patient identifier was excluded from the analysis and used solely as an index. Hospital identifier was not included as a covariate due to the high number of categories and sparse representation across hospitals. Outcome-related variables and their associated time variables were excluded from the set of covariates to avoid target leakage and ensure a clear separation between predictors and outcomes.

The final set of covariates included patient age at diagnosis, year of surgery, menopausal status, tumor size category, histological grade, number of positive lymph nodes, ER and PGR status, and indicators of adjuvant treatments (chemotherapy and hormonal therapy).

Continuous variables exhibiting skewed distributions were transformed to improve model stability. In particular, the number of positive lymph nodes was log-transformed after adding a small constant, resulting in the

variable nodes_log, to mitigate right-skewness and reduce the influence of extreme values. ER and PGR measurements were dichotomized using a threshold of 10 fmol/mg, in accordance with standard clinical practice and the project guidelines. Tumor size and grade were treated as ordinal variables to preserve their inherent ordering.

For classification models and time-to-event analyses, interaction terms between tumor size and grade were additionally explored to capture potential effect modification between these key prognostic factors. Age at diagnosis was modeled as a continuous variable, with nonlinear effects accommodated where appropriate in downstream analyses. Such derived features were intentionally excluded from the causal inference analysis in order to maintain model interpretability and avoid unnecessary complexity in the causal structure.

All covariates were preprocessed using a unified pipeline across modeling frameworks. Continuous variables were standardized using z-score normalization, ordinal variables were encoded using ordinal encoding, and nominal categorical variables were one-hot encoded. This preprocessing strategy was adopted to ensure consistent variable representation and stable estimation across linear modeling approaches, including classification, survival, and causal inference analyses.

2.3 Statistical Models

2.3.1 Classification Models

Classification models were developed to predict two binary outcomes: tumor recurrence and all-cause mortality. Logistic regression models with L2 regularization were employed due to their interpretability, robustness, and suitability for clinical risk prediction. Regularization was used to mitigate overfitting and stabilize coefficient estimates in the presence of correlated covariates.

Model development followed a structured, multi-step pipeline. Baseline models were first fitted using clinically relevant covariates. Regularization strength was subsequently tuned using grid search with stratified 5-fold cross-validation. Feature engineering was then explored by introducing derived features, including interaction terms between tumor size and histological grade, as well as a quadratic term for age. Feature selection was performed using backward sequential feature selection with cross-validation, optimizing the ROC–AUC metric. When derived features were retained, the corresponding base covariates were preserved in the model to maintain interpretability. Following feature selection, regularization strength was re-tuned on the reduced feature set.

Model performance was evaluated using ROC–AUC on a held-out test set after each modeling step, and final models were selected by balancing predictive performance, parsimony, and clinical interpretability.

For recurrence prediction, the classification threshold was not fixed at 0.5. Instead, the decision threshold was selected to prioritize sensitivity, reflecting the higher clinical cost of missing true recurrence cases compared to false positive predictions.

To complement the predictive analysis, logistic regression models were additionally fitted using the statsmodels framework on the final set of selected covariates. This inference-oriented analysis enabled estimation of odds ratios, confidence intervals, and statistical significance, which are not directly available in the predictive modeling framework.

2.3.2 Time-to-Event Models

Time-to-event analyses were conducted to investigate recurrence-free survival and overall survival while explicitly accounting for censoring and variable follow-up times. These models complement the classification analysis by providing a dynamic assessment of risk over time rather than a fixed binary outcome.

Cox proportional hazards models were employed to estimate the association between covariates and the hazard of recurrence or death. Covariates and preprocessing steps were consistent with those used in the classification models to ensure comparability across analytical frameworks. Model development followed a stepwise strategy analogous to the classification pipeline, including baseline model fitting, regularization strength tuning, exploration of derived features, and feature selection.

The proportional hazards assumption was formally assessed for all Cox models. Violations of this assumption were observed for multiple covariates in both recurrence-free survival and overall survival analyses. Several strategies were employed to mitigate these violations, including stratification on selected categorical variables and the introduction of time-dependent effects for covariates exhibiting clear time-varying behavior. These adjustments led to improved model fit and predictive performance, as assessed by concordance indices. However, some departures from proportional hazards persisted in the final models. Given that further corrective measures resulted in substantial loss of model performance or interpretability, the final specifications were retained as a pragmatic compromise between model validity, predictive accuracy, and clinical interpretability.

Model results were summarized using hazard ratios and corresponding confidence intervals. Final model specifications were selected based on a balance between goodness of fit, interpretability, and consistency with clinical knowledge.

2.3.3 Causal Inference Analysis

In addition to predictive and survival modeling, a causal inference analysis was conducted to estimate the average treatment effect (ATE) of adjuvant hormonal therapy on tumor recurrence and all-cause mortality using observational data. Unlike classification models, the objective of this analysis was not prediction but causal effect estimation under explicit assumptions.

Hormonal therapy was treated as the exposure of interest, with recurrence and mortality as outcomes. Given the non-randomized nature of treatment assignment and the relatively low proportion of treated patients (11.4%), potential confounding by indication and limited covariate overlap were explicitly considered. Adjustment covariates included baseline patient and tumor characteristics measured prior to treatment initiation, namely age at diagnosis, year of surgery, menopausal status, tumor size category, histological grade, number of positive lymph nodes, ER and PGR status, and chemotherapy.

As an initial approach, inverse probability of treatment weighting (IPTW) based on propensity scores was explored to balance treatment groups. Propensity scores were estimated using logistic regression, followed by trimming strategies to improve overlap. While aggressive trimming criteria achieved substantial covariate balance, they resulted in considerable data loss and limited population representativeness. Alternative trimming strategies preserved a larger portion of population and improved balance but did not fully eliminate residual imbalance for clinically relevant covariates. Given the sensitivity of IPTW-based estimates to limited overlap and remaining imbalance, this approach was not retained as the primary causal analysis.

Consequently, a regression-based outcome modeling approach was adopted to estimate the treatment effect of hormonal therapy. Linear modeling frameworks were used to adjust for the selected covariates. In contrast to the predictive analyses, no interaction terms or nonlinear transformations were included, prioritizing interpretability of the estimated treatment effect.

All causal effect estimates were summarized as point estimates of the average treatment effect on the risk scale, with uncertainty quantified using bootstrap-based confidence intervals, and interpreted under the standard assumptions of conditional exchangeability, positivity, and consistency. While residual unmeasured confounding cannot be excluded, this approach provides a transparent and clinically interpretable estimate of the effect of hormonal therapy within the limits of the available observational data.

3. Results

3.1 Classification Results

Final classification models for recurrence and mortality are summarized in Table 2. Regularized logistic regression achieved stable discriminative performance across outcomes, with test ROC–AUC values in the moderate-to-good range.

Table 2. Final classification models and predictive performance.

Outcome	Final predictors	Regularization (C)	Test ROC–AUC	Threshold	Test Recall
Recurrence	['num__nodes_log' 'age__age' 'cat__hormon_1' 'cat__chemo_1' 'ord__size' 'ord__grade']	2.15	0.744	0.38	0.847
Mortality	['num__year' 'num__nodes_log' 'age__age' 'age__age^2' 'cat__meno_1' 'cat__er_pos_1' 'cat__pgr_pos_1' 'cat__hormon_1' 'cat__chemo_1' 'ord__size' 'ord__grade' 'ord__size grade']	1.0	0.82	0.31	0.855

Inference-oriented logistic regression analyses revealed consistent prognostic patterns across outcomes. Tumor-related factors, including lymph node involvement, tumor grade, and tumor size, showed the strongest and most stable associations with both recurrence and mortality risk. In contrast, treatment-related variables were associated with reduced risk, with effects more pronounced for recurrence than for mortality. Age exhibited a weaker and outcome-dependent effect, including a nonlinear association with mortality. Several additional covariates did not reach statistical significance in multivariable models, which is expected in the presence of correlated predictors and

interaction terms. Overall, the observed associations align with established clinical knowledge and support the validity of the final model specifications.

3.2 Time-to-Event Results

Results from the time-to-event analysis are summarized in Table 3. Overall, survival models identified prognostic patterns consistent with the classification analysis, while additionally capturing time-dependent effects.

Table 3. Final Cox models for recurrence-free and overall survival

Outcome	Stratification by	Model features (key associations in bold, HR ≠ 1) *	Concordance index
Recurrence-free survival	cat__pgr_pos_1	['**year_x_logtime**' '**num__year**' '**num__nodes_log**' '**age__age**' '**age__age^2**' 'cat__meno_1' 'cat__er_pos_1' '**cat__hormon_1**' '**cat__chemo_1**' '**ord__size**' '**ord__grade**' '**ord__size grade**']	0.74
Overall survival	cat__pgr_pos_1, cat__er_pos_1	['num__year' '**num__nodes_log**' 'age__age' '**age__age^2**' 'cat__meno_1' '**cat__hormon_1**' '**cat__chemo_1**' '**ord__size**' '**ord__grade**' 'ord__size grade']	0.72

* Bold text indicates covariates with hazard ratios different from 1 in the final model.

Tumor-related factors, including lymph node involvement, tumor grade, and tumor size, were consistently associated with increased hazards of recurrence and death. Treatment-related variables showed protective associations, with effects generally more pronounced for recurrence-free survival than for overall survival.

For recurrence-free survival, a time-dependent effect of year of surgery was observed, indicating that the protective effect of more recent diagnosis attenuates over follow-up time.

To accommodate heterogeneity in baseline risk, models were stratified by progesterone receptor status for recurrence-free survival and by combined estrogen and progesterone receptor status for overall survival. Despite diagnostic adjustments, some departures from the proportional hazards assumption remained in the final models. These were retained as a pragmatic trade-off between model adequacy, predictive performance, and interpretability.

3.3 Causal Inference Results

Results of the causal inference analysis are summarized in Table 4. Overall, hormonal therapy was associated with a reduction in the risk of both recurrence and all-cause mortality; however, estimated effects were modest and accompanied by substantial uncertainty.

Table 4. Estimated average treatment effect (ATE) of hormonal therapy

Outcome	ATE (risk difference)	95% CI
Recurrence	-0.15	from -0.20 to -0.09
Mortality	-0.10	from -0.15 to -0.05

While point estimates suggested a protective effect of hormonal therapy, confidence intervals included values close to zero, indicating limited statistical precision. These findings are consistent with the limited overlap between treated and untreated patients and the relatively small proportion of patients receiving hormonal therapy. Consequently, causal effect estimates should be interpreted cautiously, as indicative rather than definitive.

4. Discussion and Conclusions

This study applied complementary statistical modeling frameworks to investigate prognosis and treatment effects in breast cancer using observational cohort data from the Rotterdam Breast Cancer Study. By combining classification models, time-to-event analysis, and causal inference, we obtained a coherent and multifaceted view of factors associated with recurrence and mortality, while explicitly addressing the limitations inherent to non-randomized data.

Across classification and survival analyses, tumor-related characteristics (particularly lymph node involvement, tumor grade, and tumor size) emerged as the most consistent and robust prognostic factors for both recurrence and all-cause mortality. These findings are well aligned with established clinical evidence identifying tumor burden and disease aggressiveness as key determinants of long-term outcomes in breast cancer [1,3]. Treatment-related variables, including hormonal therapy and chemotherapy, were associated with reduced risk across modeling frameworks, with stronger protective effects observed for recurrence-related outcomes than for mortality.

This pattern is clinically plausible, as adjuvant therapies primarily aim to reduce disease recurrence rather than competing causes of death [2].

Time-to-event analyses further enriched these findings by capturing the dynamic nature of risk over follow-up time. The observed time-dependent effect of year of surgery suggests improvements in early outcomes for patients treated in more recent periods, likely reflecting advances in diagnostics and treatment strategies. At the same time, residual departures from the proportional hazards assumption highlight the complexity of modeling long-term survival in heterogeneous clinical populations, and underscore the need for pragmatic trade-offs between model assumptions, interpretability, and predictive adequacy.

Causal inference analysis provided a more conservative perspective on the effect of hormonal therapy. While outcome regression suggested a protective average treatment effect on both recurrence and mortality, the estimates were accompanied by substantial uncertainty. This reflects the limited overlap between treated and untreated patients, the relatively low prevalence of hormonal therapy in the cohort, and the potential for residual confounding by indication (challenges that are well documented in observational oncology research) [5]. As in any observational study, residual unmeasured confounding cannot be excluded, particularly for clinical factors influencing treatment assignment that are not fully captured by the available covariates [4]. The contrast between stronger predictive associations and more modest causal estimates underscores the fundamental distinction between prediction and causal interpretation.

Several limitations should be acknowledged. First, treatment assignment was not randomized, and unmeasured confounding cannot be ruled out despite careful covariate adjustment. Second, some modeling assumptions (most notably proportional hazards) were only partially satisfied. Finally, the cohort reflects historical treatment practices, which may limit direct generalizability to contemporary clinical settings.

In conclusion, this project demonstrates how different statistical frameworks can provide complementary insights into breast cancer prognosis and treatment effects. Regularized classification models and survival analysis offer robust tools for risk stratification, while causal inference methods yield more cautious but interpretable estimates of treatment effects. Together, these approaches highlight both the opportunities and the limitations of observational healthcare data, and emphasize the importance of aligning modeling choices with clearly defined analytical objectives.

References

1. Pedersen, R.N.; Cronin-Fenton, D.; Heide-Jørgensen, U.; et al. The incidence of breast cancer recurrence 10–32 years after primary diagnosis. J. Natl. Cancer Inst. 2022
2. Early Breast Cancer Trialists' Collaborative Group. Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. Lancet 2011
3. Royston, P.; Altman, D.G. External validation of a Cox prognostic model: principles and methods. BMC Med. Res. Methodol. 2013
4. Hernán, M.A.; Robins, J.M. Causal Inference: What If. Chapman & Hall/CRC: Boca Raton, FL, USA, 2020
5. Holmberg, L.; Anderson, H.; HABITS steering and data monitoring committees. Increased risk of recurrence after hormone replacement therapy in breast cancer survivors. J. Natl. Cancer Inst. 2008