STATISTICAL MODELS FOR HEALTHCARE DATA
2025-2026 EXAM PROJECT

Breast cancer is a heterogeneous disease whose prognosis depends on a wide range of patient, tumor, treatment, and healthcare system factors. Clinical management decisions, such as whether to administer adjuvant chemotherapy or hormonal therapy, are informed by risk stratification models that estimate the likelihood of recurrence and mortality. Yet despite decades of research, substantial uncertainty remains regarding how individual prognostic factors combine to shape long-term outcomes, and how treatment effects can be reliably estimated in observational settings where therapies are not randomly assigned.

In this project, you will use data from the **Rotterdam Breast Cancer Study** [1], a well-known cohort with extended follow-up, to investigate prognosis and treatment effects using the analytical frameworks addressed during the course. Specifically, you are asked to explore the impact of patients' characteristics with respect to the following outcomes and modeling pipelines:

1. **A classification model** to predict tumor recurrence.
2. **A classification model** to predict long-term mortality.
3. **Time-to-event models** to examine recurrence-free survival and overall survival.
4. **A causal inference analysis** to estimate the effect of adjuvant hormonal therapy on outcomes.

Particular attention shall be paid to the interpretation of the final results.

The exam is deemed submitted **only when the following materials have been provided**: (1) the **project code**, which must be reproducible and properly documented, and (2) **a written report of up to six pages**, to be handed in no later than one week before the scheduled exam session. During the exam, you must deliver an **oral presentation** of your project (7 minutes), after which you will be asked **follow-up questions**. See the Deadline section for details.

The report shall be structured as follows.

Part 1: Introduction

Write a brief overview of your research proposal, addressing the following:

- Research Question: Define clear, focused research questions that your study will address.
- Hypothesis: Based on prior findings, state your hypotheses. For example, how do you expect different treatments to affect mortality? Which confounders do you expect?
- Rationale: Explain why this research question is important. Refer to existing literature on breast cancer therapies, including but not limited to [2-4].

Part 2: Materials and Methods

Develop a basic outline for the research methods you would use, considering the following components:

- Study Population and Sample: Describe the characteristics of your study participants (e.g., number of participants, key demographics), stratified per treatment status, and for every stratification you aim to analyze (e.g. tumor grade).
- Covariates: List relevant covariates you would control for in your analysis with their distributions (descriptive statistics).
- Outcome Measurement: Describe the main outcomes you will examine.
- Statistical Models: Specify the type of models and the type of causal inference approach(es) you used, providing rationales for and details of your choices.

Part 3: Results

In the Results section of your report, summarize the key findings based on the dataset provided.

Include coefficient values and confidence intervals for the different models, where available, and all relevant metrics. Include any figure and table you deem appropriate, with clear captions and cross-reference in the text.

tables and plots should be referenced

Part 4: Discussion and Conclusions

Write a brief discussion of your research findings and how they might contribute to public health, clinical practice, and breast-cancer management strategies. Consider how your findings could inform future treatment guidelines for breast cancer patients. Discuss the role of confounding and how it was addressed in the analysis. Highlight limitations, such as potential unmeasured confounding, and their implications for causal inference.

Grading Criteria:

1. Clarity and focus of research question and hypothesis
2. Coherence between python code and the methodology and results described
3. Relevance and depth of methodology address all the questions based on the methods we've seen
4. Appropriateness and rigor of statistical analysis approach
5. Quality of the presentation
6. Insightfulness of discussion

Dataset Variable Descriptions:

The dataset comprises 2982 primary breast cancer patients whose records were included in the Rotterdam tumour bank. The following variables have been collected:

Baseline clinical / pathological variables

- pid: patient identifier
- age: age at diagnosis (years)
- year: year of surgery
- hospital_id: id of the hospital where the patient was treated (1-25)
- meno: menopausal status (0= premenopausal, 1= postmenopausal)
- nodes: number of positive lymph nodes
- size: tumor size category (<=20, 20-50, >50 mm)
- grade: histological grade (1–3, higher grade indicates more severe disease)
- er: estrogen receptors (fmol/l)
- pgr: progesterone receptors (fmol/l)
- hormon: adjuvant hormonal therapy (1=yes, 0=no)
- chemo: adjuvant chemotherapy (1=yes, 0=no)

Time-to-event variables

- rtime: time from surgery to first recurrence or censoring (days)
- recur: recurrence indicator (1=recurrence, 0=censored)
- dtime: time from surgery to death or censoring (days)
- death: death indicator (1=death, 0=censored)

ER and PGR are central biomarkers in breast cancer biology and strongly influence prognosis and treatment decisions. Tumors that are ER-positive and/or PGR-positive rely (partially) on estrogen and progesterone signaling for growth. Values > 10fmol/mg can be considered ER/PGR positive.

These patients have undergone different types of treatment:

- Hormonal (Endocrine) Therapy (*hormon=1)*, which blocks estrogen signaling or reduces estrogen production, slowing growth of hormone-dependent cancer cells. Examples include *Tamoxifen* and *Aromatase inhibitors*.
- Chemotherapy (*chemo=1),* which uses cytotoxic drugs that kill rapidly dividing cancer cells, reducing relapse risk even in hormone-insensitive tumors.

There are subjects who have died without recurrence, but whose death time is greater than the censoring time for recurrence. A common way that this happens is that a death date is updated in the health record sometime after the research study ended, and said value is then picked up when a study data set is created.

Deadline:
You must register for the exam session in which you intend to take the exam by following the UniSR Exam Registration System guidelines and deadlines. Before the selected exam date, you are required to submit your materials by the specified deadline. Exact dates and times for each session are listed below:

Exam: Monday, **February 2nd, 2026** → Deadline: Friday, **January 23rd, 2026**
Exam: Friday, **February 13th, 2026** → Deadline: Friday, **February 6th, 2026**
Exam: Thursday, **April 9th, 2026** → Deadline: Friday, **March 27th, 2026**
Exam: Monday, **June 22nd, 2026** → Deadline: Friday, **June 12th, 2026**
Exam: Monday, **July 6th, 2026** → Deadline: Friday, **June 26th, 2026**
Exam: Wednesday, **September 2nd, 2026** → Deadline: Friday, **August 21st, 2026**
Exam: Wednesday, **September 16th, 2026** → Deadline: Friday, **September 4th, 2026**

The submission must consist of a single zipped folder — named "surname_studentID" — containing both the project code and the report. The upload link will be provided via an announcemet on the Blackboard platform prior to each exam session.

The report must not exceed six pages, and the project code must be well organized, fully commented, and reproducible; a README file should be included when necessary. On the day of the exam, students will deliver a 7-minute oral presentation of their project, followed by a short Q&A session.

REFERENCES

[1] Royston, Patrick, and Douglas G. Altman. "External validation of a Cox prognostic model: principles and methods." *BMC Medical Research Methodology* 13.1 (2013): 33.

[2] Holmberg, Lars, et al. "Increased risk of recurrence after hormone replacement therapy in breast cancer survivors." *Journal of the National Cancer Institute* 100.7 (2008): 475-482.

[3] 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." *The Lancet* 378.9793 (2011): 771-784.

[4] Pedersen, Rikke Nørgaard, et al. "The incidence of breast cancer recurrence 10-32 years after primary diagnosis." *JNCI: Journal of the National Cancer Institute* 114.3 (2022): 391-399.