Survival Analysis on the SEER Breast Cancer Dataset

When Does Machine Learning Become Useful?

Lirida Papallazi

Università degli Studi di Milano

2020/2021

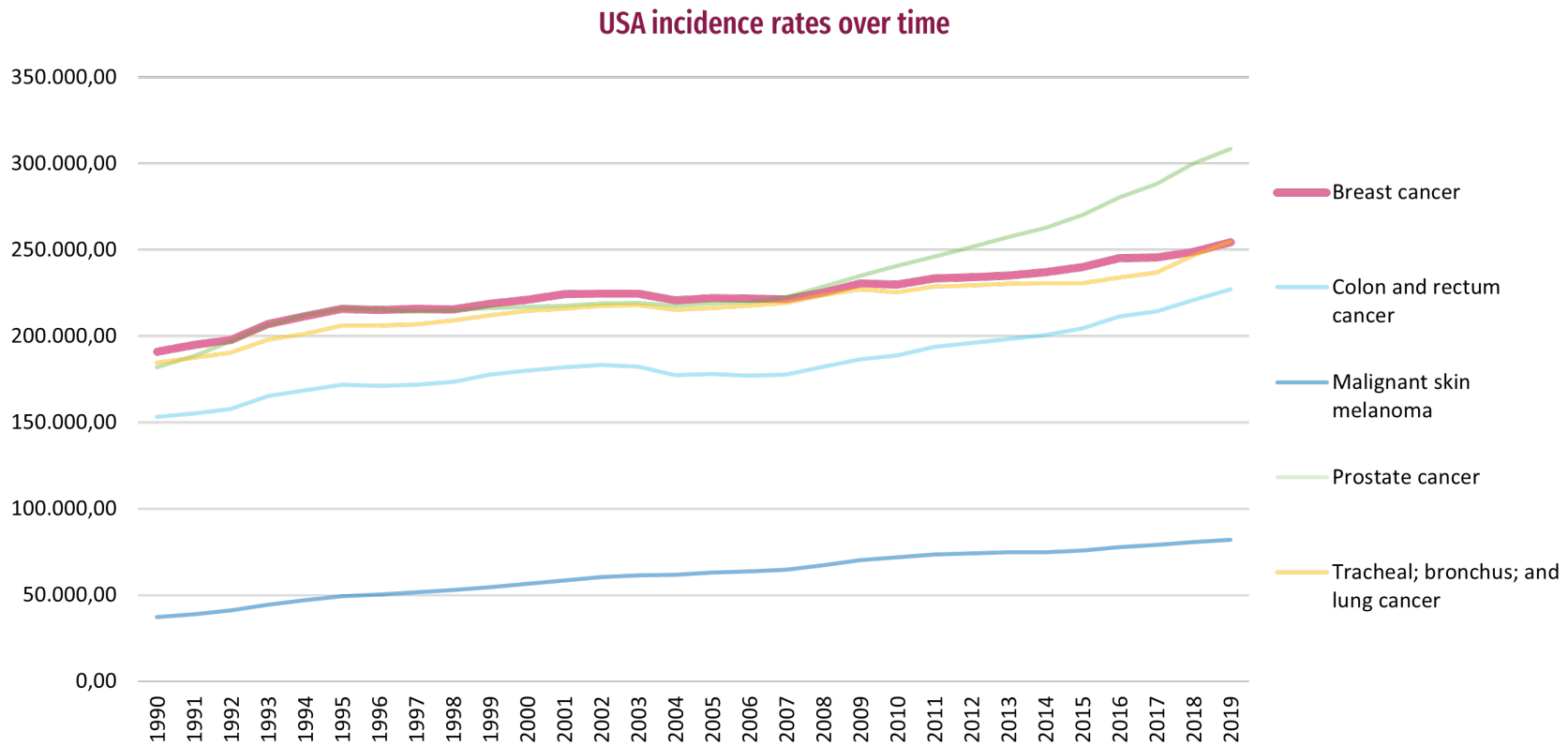
UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE POLITICHE,
ECONOMICHE E SOCIALI

Outline

The problem

Starting from the first three decades of the 20th century, cancer has begun to represent one of the major causes of death all over the world.[3]

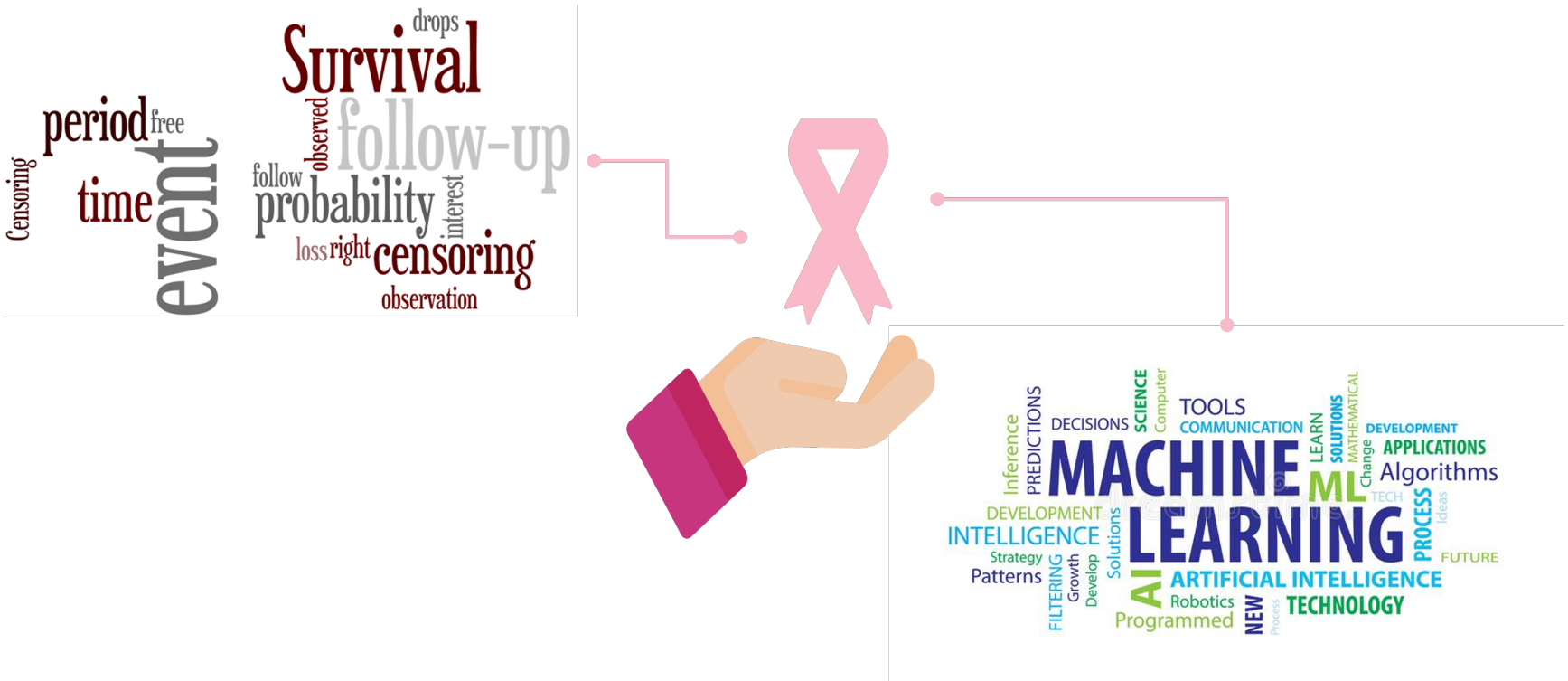


The problem (cont.)

The impact of this disease and the availability of a very large literature on the subject have led to the decision to focus on breast cancer.

The goal

The ultimate purpose of the thesis is to understand if "modern" machine learning algorithms bring an actual added value compared to the classic survival analysis models, in predicting the survival of patients diagnosed with breast cancer.



Features of Survival Analysis

Survival analysis is a collection of statistical procedures for data analysis where the outcome variable of interest is time until an event occurs.

01

Censoring

At the end of the clinical study some of the subjects have not experienced the event of interest or they are lost to follow up.

02

Survival function

This function represents the probability that the event of interest has not occurred at time *t:*

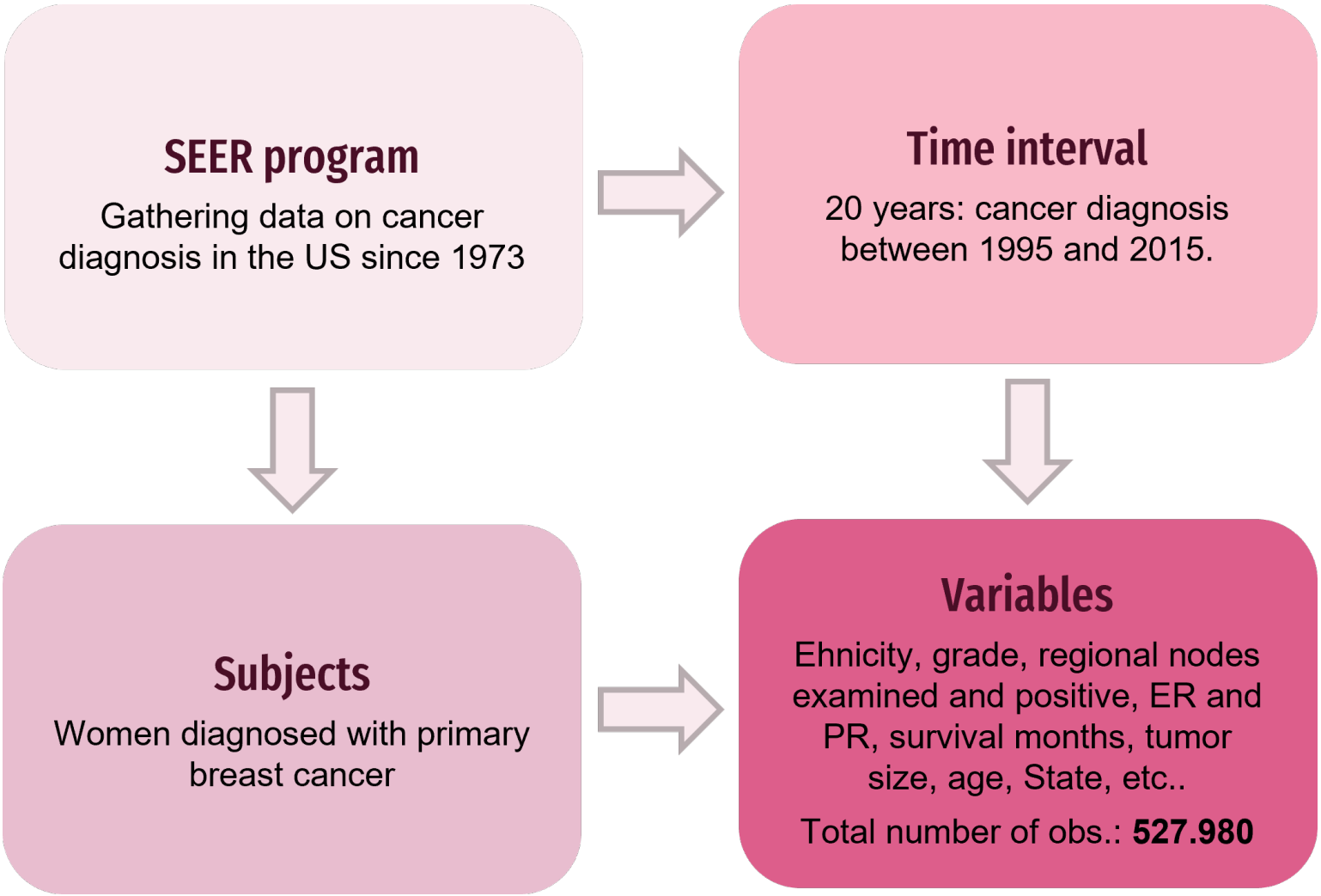
$S(t) = Pr(T>t)$

03

Hazard function

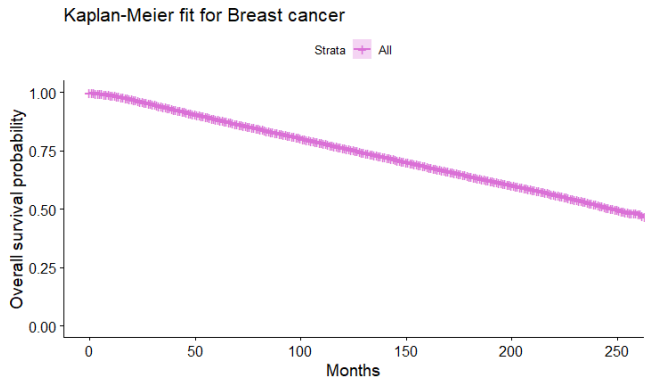
The instantaneous probability that a subject will verify the event of interest

Building the dataset



Analysis: Part I

Kaplan-Meier Non-parametric statistic which studies the survival times of members of a group.



Log-Rank test Method used to compare the survival curves of two or more groups.

```
> survdiff(Surv(survival_days, status) ~ grade, data=training_20y)
Call:
survdiff(formula = Surv(survival_days, status) ~ grade, data = training_20y)

                                                N Observed Expected (O-E)^2/E (O-E)^2/V
grade=Moderately differentiated; Grade II    152818    30989    33573     198.9     350.4
grade=Poorly differentiated; Grade III       123865    33775    26174    2207.3    3332.6
grade=Undifferentiated; anaplastic; Grade IV   3841     1377     1170      36.5      37.2
grade=Well differentiated; Grade I            76652    11767    16990    1605.9    2059.3

 Chisq= 4060  on 3 degrees of freedom, p= <2e-16
```

PH and Cox model Semi-parametric model that determines the relationship between the survival time and the impact of the covariates on the survival distribution.[1]

```
> prop_hazards6 <- cox.zph(fit_breast_cox6)
> print(prop_hazards6)
                          chisq df      p
ethnicity                   460  2 <2e-16
grade                      4261  3 <2e-16
estrogen_receptor          5033  2 <2e-16
progesteron_receptor       3666  2 <2e-16
ns(age, 3)                 4697  3 <2e-16
regional_nodes_examined     367  1 <2e-16
regional_nodes_positive     562  1 <2e-16
ns(tumor_size, 3)          2758  3 <2e-16
GLOBAL                    10824 17 <2e-16
```

Analysis: Part I (cont.)

AFT models Models that can be used when the PH assumption is violated and they assume that the covariates accelerate or decelerate the hazard rate by some constant.

AIC and BIC results for AFT models

Model	AIC	BIC
Exp	4.298	4.394
Weibull	4.274	4.374
Log-normal	4.318	4.419
Log-logistic	4.261	4.362

Model Validation Validation of the model on increasing sizes of the original training set.

C-index for increasing sample size

Sample size	C-index
20%	0.748 57
30%	0.749 83
40%	0.748 73
50%	0.751 36
60%	0.744 96
70%	0.749 33
80%	0.749 16
90%	0.749 99
100%	0.749 77

Analysis: Part II

In the second part we instead applied 5 machine learning algorithms implementing the following steps:

1	2	3	4	5
Split the original dataset into two equal parts: training set and external validation, each consisting of over **260.000 observations**	Train the machine learning algorithms on increasing sample sizes of the training set drawn at random from the training set	Hyperparameters tuning of the models via cross validation	Evaluation of the models on the external validation set via computation of the C-index	Comparison of the results obtained in step (4) with the Cox model fitted on the external validation

Conclusion

Observing the results, it is possible to state that there is a real added value brought by machine learning algorithms compared to the more classic Cox model.

Harrell's C-index for the models implemented

Model	0.1%	0.5%	1%	5%
Observations	360	1.800	3.600	18.000
Cox	68.81%	69.37%	73.62%	74.10%
Random Forest	70.31%	72.90%	73.01%	74.20%
Gradient Boosted	66.77%	72.00%	72.04%	72.72%
SVM	67.86%	68.04%	68.10%	68.22%
CoxTime	97.71%	98.31%	98.40%	99.27%
DeepSurv	87.23%	88.00%	88.63%	91.76%
DeepSurv (PyTorch)	60.10%	62.45%	65.05%	66.22%

Main issues

Numerous problems had to be addressed during our analysis:

Novelty of the topics The lack of knowledge in the epidemiological field required a phase of preparation and understanding of the topics and gathering of the correct information.

SEER database and variable selection Understanding of the functioning of the SEER program and of the classifications used and selection of the variables of interest.[2]

Tools The use of both R and Python for the analysis required different pre-processing techniques and methods of communication between the results obtained by the two languages.

Main issues (cont.)

Numerous problems had to be addressed during our analysis:

Packages CoxTime and DeepSurv have been evaluated with Antolini's time dependent concordance index and not Harrell's C-index.[4]

Hyperparameter tuning The ability to optimize machine learning models when dataset size becomes important requires more computing and memory power.

Future projects

This thesis is intended only to be a starting point for future and more in-depth analyzes:

Optimization	Other variables?	mlr3	Metrics
Further optimization of the algorithms implemented	Inclusion of variables describing the behaviours of the patients and other pathologies	Implementation and comparison with the results obtained using the mlr3 package still under development	Extension of the IPA and concordance index also to the machine learning algorithms

Thank you for your attention!

References


D.F. Moore.
Applied Survival Analysis Using R.


SEER.
Variable and Recode Definitions.


breastcancer.org.
U.S. Breast Cancer Statistics.


L. Antolini, P. Boracchi, E. Biganzoli.
A time-dependent discrimination index for survival data.