NKR_Synthetic_Data dataset – Opportunities of data analysis and machine learning (ML)

By Grzegorz Furdyn

March 2022

Objective

This report has two objectives:

1. Demonstrate examples of exploratory data analysis of “NKR_synthetische_data”, showing influence of various features on patient’s survival
2. Present opportunities of future machine learning model applied to “NKR_synthetische_data”

Data analysis

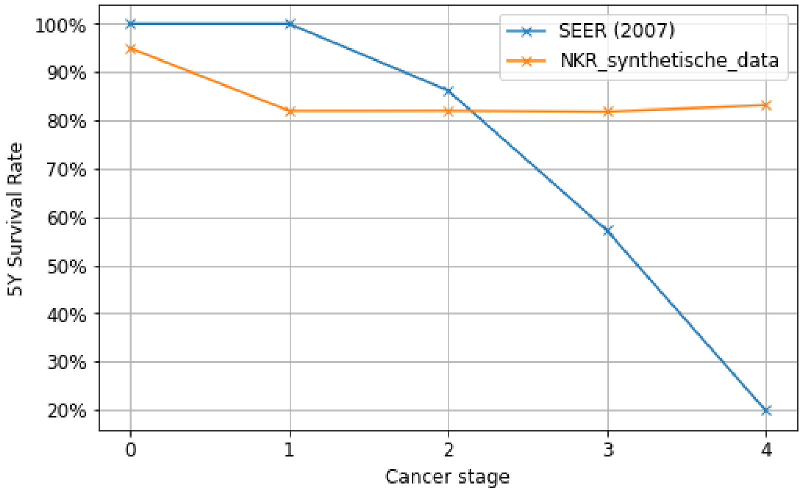
The example results shown below are based on similar work of SEER (2007). The SEER monograph presents results based on 300 000 cases of breast cancer from the US Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI). It focuses on the influence of factors such as tumor grade, size, morphology, lymph node status and many others on patients’ survival.

In this report, I have mimicked some of the SEER analyses, using “NKR_synthetische_data”, and compared them with SEER results.
“NKR_synthetische_data” consists of 60 000 breast cancer cases, described by 42 patient and medical parameters. Individual outcome is described by vit_stat (vital status of the patient) and time after detection - vit_stat_int. Based on these features, I derived parameter “survival_5Y” (5 years survival rate). I used a subset of “NKR_synthetische_data”, with year 2010 cohort and only females.

Results

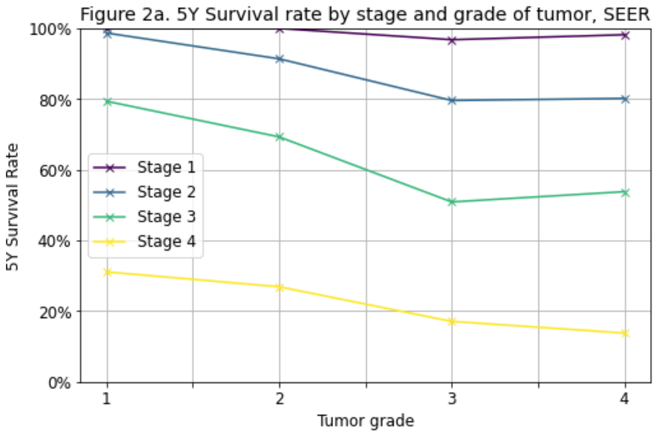
It can be expected that survival rates would vary by cancer stage (Figure 1). For direct comparison with SEER, stage has been simplified from number with letter (e.g. 1A, 1B etc.) to number only.

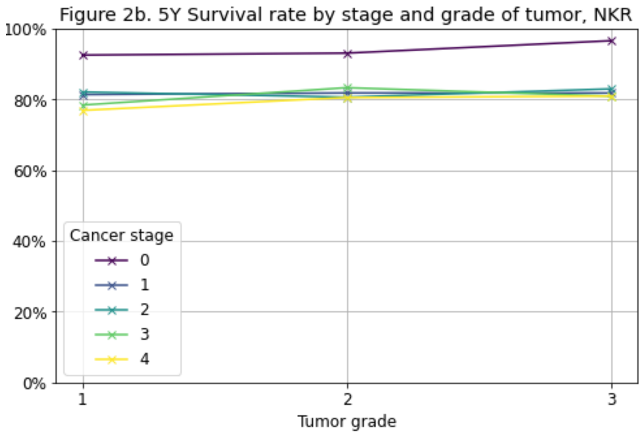
Figure 1. 5Y Survival rate as function of tumor stage at detection, SEER vs NKR



The character of the relationship seen in NKR data is very different from SEER. It likely results from the fact that data are synthetic and not realistic. It should be noted that survival rates in SEER are relative survival rates (RSR), while the rates I calculated are absolute, because of missing information necessary to recalculate as RSR. This should have little impact on the character of the relationships and will be rectified later.

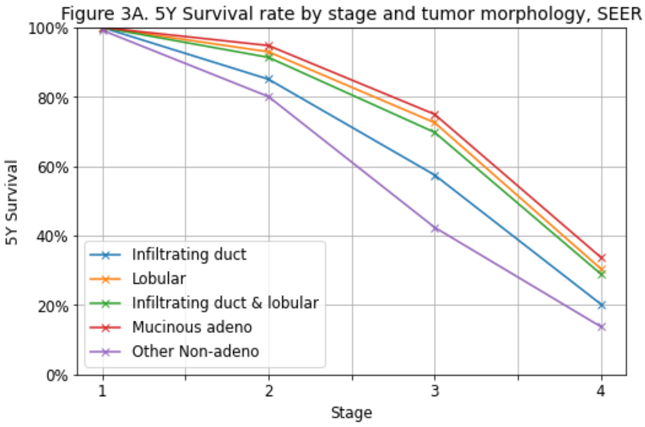
As next, 5Y survival rates by stage and grade are shown (Figure 2a, b)

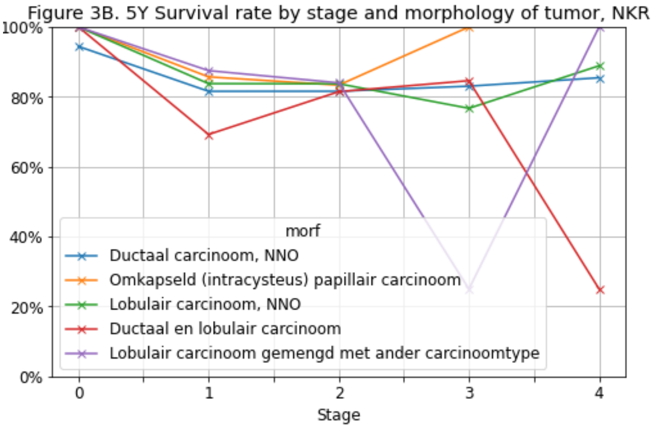




Again, the character of the relationship is different in the two datasets, in SEER being much more clear and pronounced.

As next, survival rates by stage and tumor morphology are shown (Figure 3a, b). Out of many more types of tumors in each dataset, 5 most frequent per set were selected. They is an overlap in the sets but they are not exactly the same, perhaps because of different typologies used.





Similarly, relationship in SEER dataset is more clear.

Another interesting set of features includes tumor size and the number of positive lymph nodes (Figure 4a, 4b). Following the example of SEER, the number of positive nodes was grouped in three groups: negative (0 positive nodes), 1-3 positive nodes and 4+ positive nodes. For this and following correlations, no data in tabular form are present in SEER monograph, so graphs are directly copied.

Figure 4a. 5Y Survival rate by tumor size and number of nodes, SEER

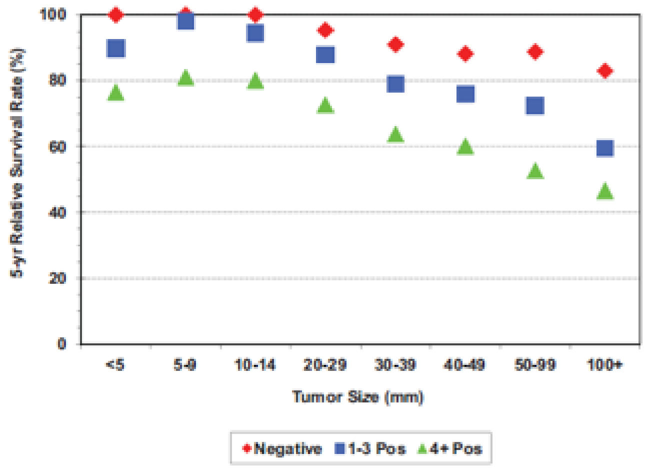
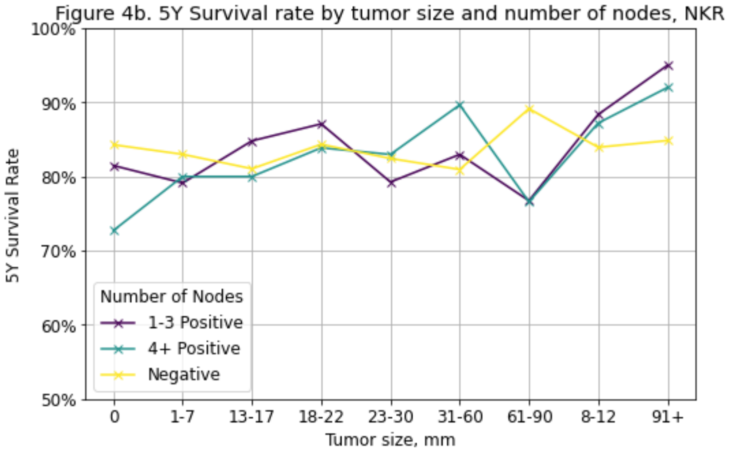


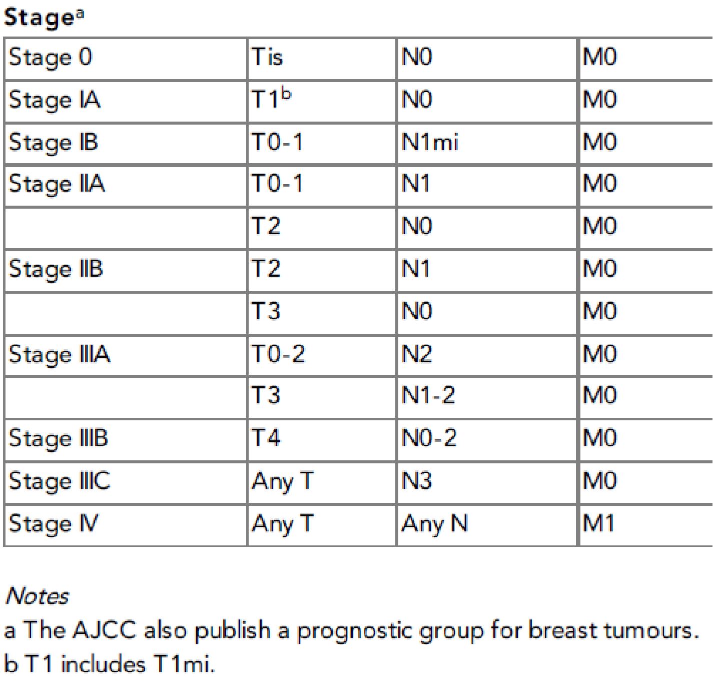
Figure 4b. 5Y Survival rate by tumor size and number of nodes, NKR



Again, character of correlations as in SEER is not observed in NKR data.

A more precise view on patients chance of survival over time can be given using the survival rates by year after diagnosis (Figure 5a,b). SEER presents these data by cancer stage identified by T,N,M status. Data from NKR are presented by stage. The relationship can be established using Table 1 (source: Richtlijnendatabase Borstkanker, Federatie Medisch Specialisten):

Table 1. Hystopathological Grading according to Nottingham Histological Score



Stage[a]

Stage 0	Tis	N0	M0
Stage IA	T1[b]	N0	M0
Stage IB	T0-1	N1mi	M0
Stage IIA	T0-1	N1	M0
	T2	N0	M0
Stage IIB	T2	N1	M0
	T3	N0	M0
Stage IIIA	T0-2	N2	M0
	T3	N1-2	M0
Stage IIIB	T4	N0-2	M0
Stage IIIC	Any T	N3	M0
Stage IV	Any T	Any N	M1

Notes

a The AJCC also publish a prognostic group for breast tumours.

b T1 includes T1mi.

Figure 5a. Relative survival rate by combinations of T, N and M, SEER

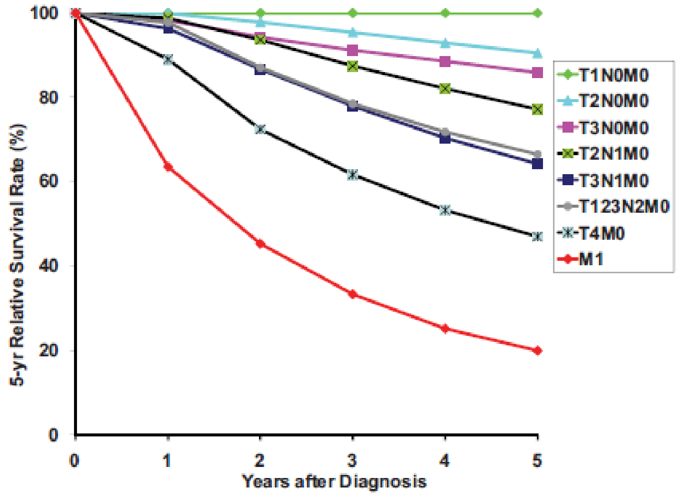
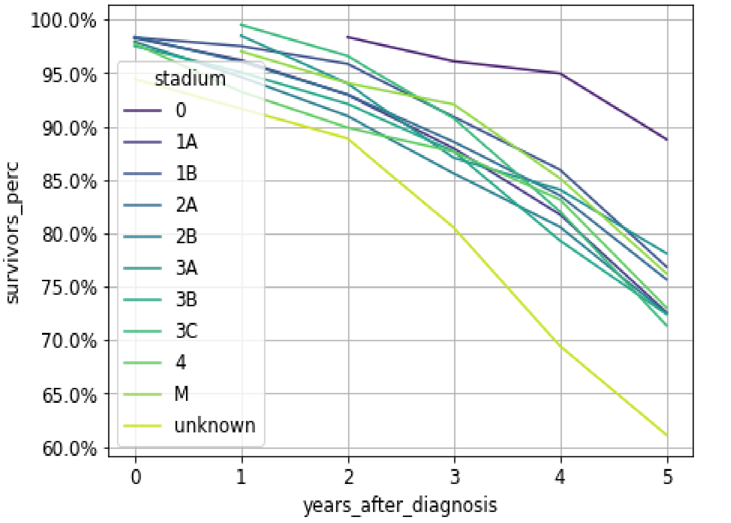
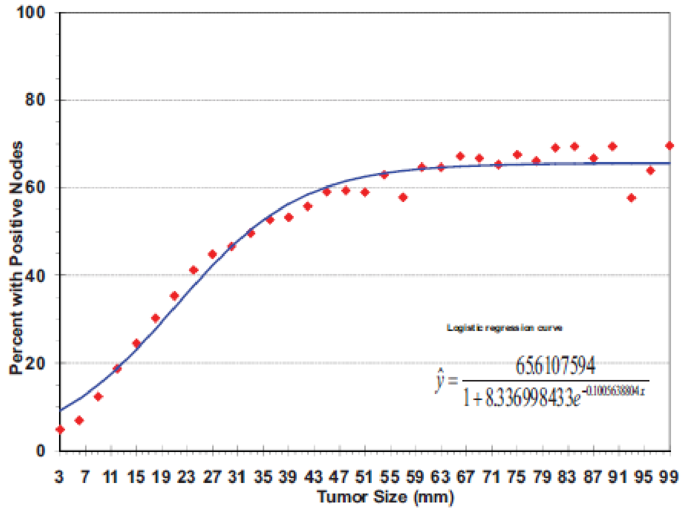


Figure 5b. Survival rate by stadium, as function of time, NKR



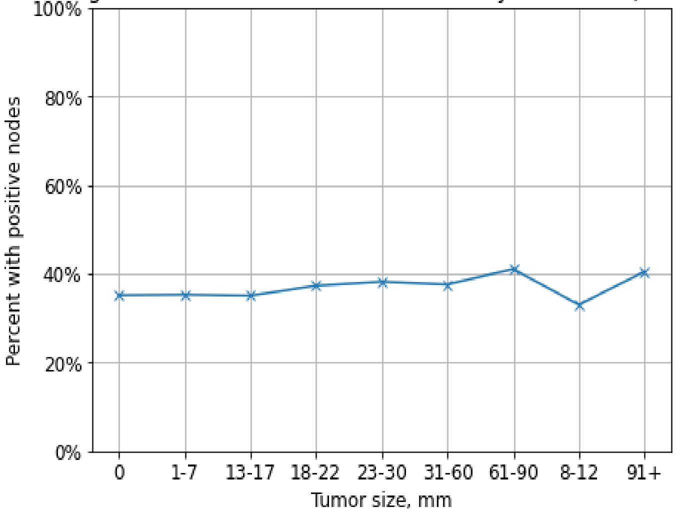
Finally, one can not only view the influence of various features on the target feature - survival rate, but also identify correlations between features. As an example, correlation between tumor size and existence of positive lymph nodes can be seen (Figure 6a,b)

Figure 6a. Existence of Positive Nodes by Tumor Size, SEER



$$\hat{y} = \frac{65.6107594}{1 + 8.336998433e^{-0.1005638804x}}$$

Figure 6b. Existence of Positive Nodes by Tumor Size, NKR



Conclusions

As could be expected, the results obtained using NKR synthetic data are very different from SEER results, which shows that the synthetic data do not reflect clinical reality. However, the tools developed in this work are ready to be used with real data to deliver realistic insights.
With real data, correlation between any features could be investigated, and the results presented in a variety of manners, in tabular and graphical form, including interactive dashboards.

Opportunity of a Machine Learning model

The NKR data present an excellent opportunities to build a machine learning model to predict patient's survival, given a set of clinical features together with patient's age (and gender, if also applied to men). The model could also be used to predict some clinical features based on others.

A draft model has been developed, using several ML techniques such as Logistic Regression and K Nearest Neighbours. However, the results in terms of accuracy, specificity and sensitivity are much below expectations so they are not shown here. This is of course due to the artificial character of the data, in which the expected correlations are not present. For the model to be optimized and validated, a set of realistic data is required.

Models of this kind have been described in literature (for overview see Li J, 2021)
With rapid development of modelling techniques, even better results of such work can be expected, and the subject is extremely interesting to explore further.

Literature

Ries LAG, Young JL, Keel GE, Eisner MP, Lin YD, Horner M-J (editors). SEER Survival Monograph: Cancer Survival Among Adults: U.S. SEER Program, 1988-2001, Patient and Tumor Characteristics. National Cancer Institute, SEER Program, NIH Pub. No. 07-6215, Bethesda, MD, 2007. Chapter 7. Cancer of the Female Breast, by Lynn A. Gloeckler Ries and Milton P. Eisner

https://seer.cancer.gov/archive/publications/survival/index.html

Richtlijnendatabase Borstkanker, **Federatie Medisch Specialisten):**
https://richtlijnendatabase.nl/richtlijn/borstkanker/tnm_8.html

Li J, Zhou Z, Dong J, Fu Y, Li Y, Luan Z, et al. (2021) Predicting breast cancer 5-year survival using machine learning: A systematic review. PLoSONE 16(4): e0250370).
https://doi.org/10.1371/journal.pone.0250370