iKNL Synthetic Data

Update May 2022
Grzegorz Furdyn

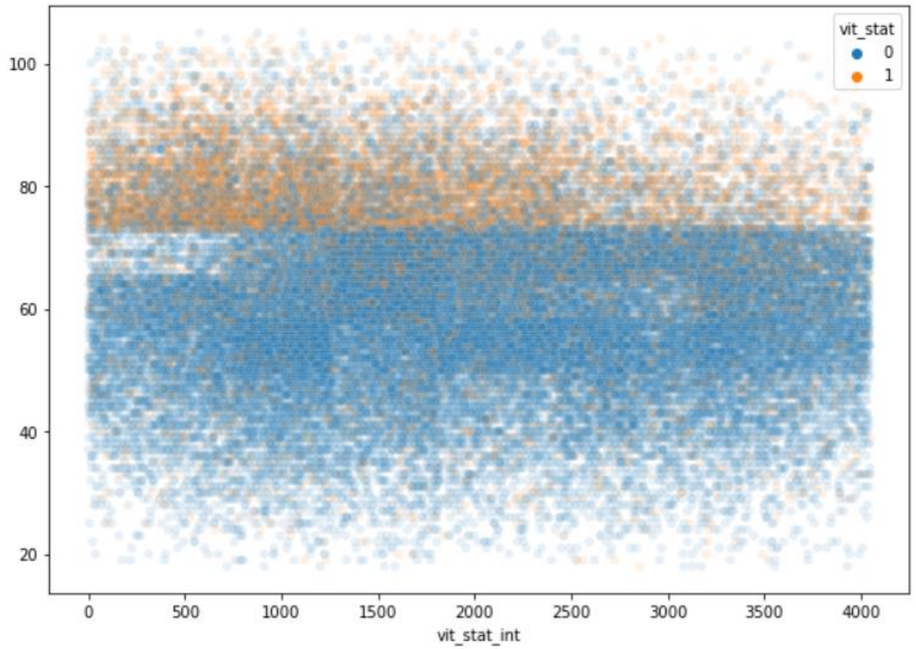
- Contents
  - Overall view of the datasets – original vs revised
  - Cancer data analysis – KNR vs SEER
  - Modeling: vital status models vs 5Y survival models
  - EDA – insights from KNR data

Overall view of the data

Original vs revised data set – overall view

hue = vital status (vit_stat)

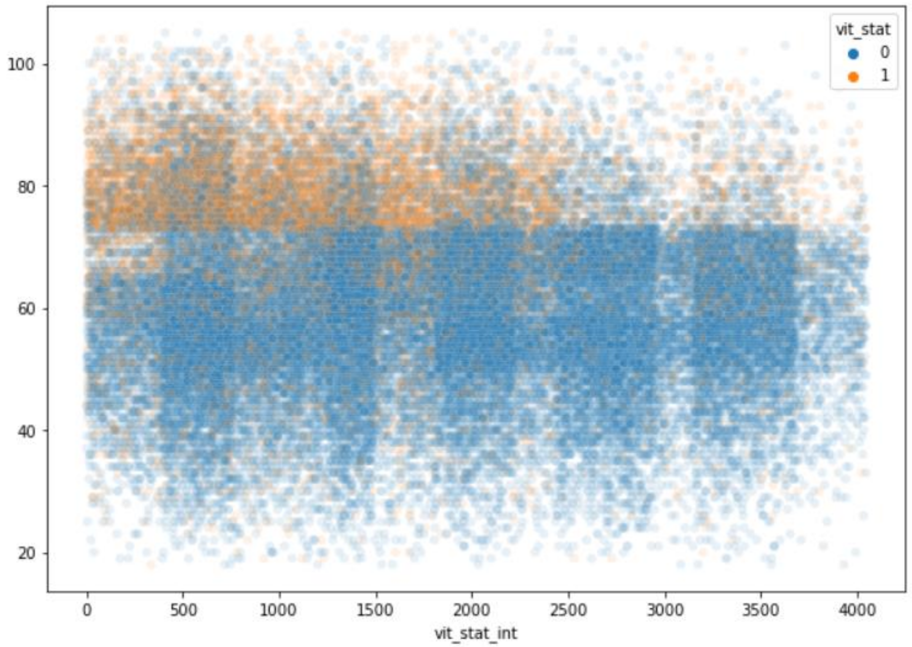
Original set



Age (leeft)

Time since incidence (vit_stat_int)

Revised set



Age (leeft)

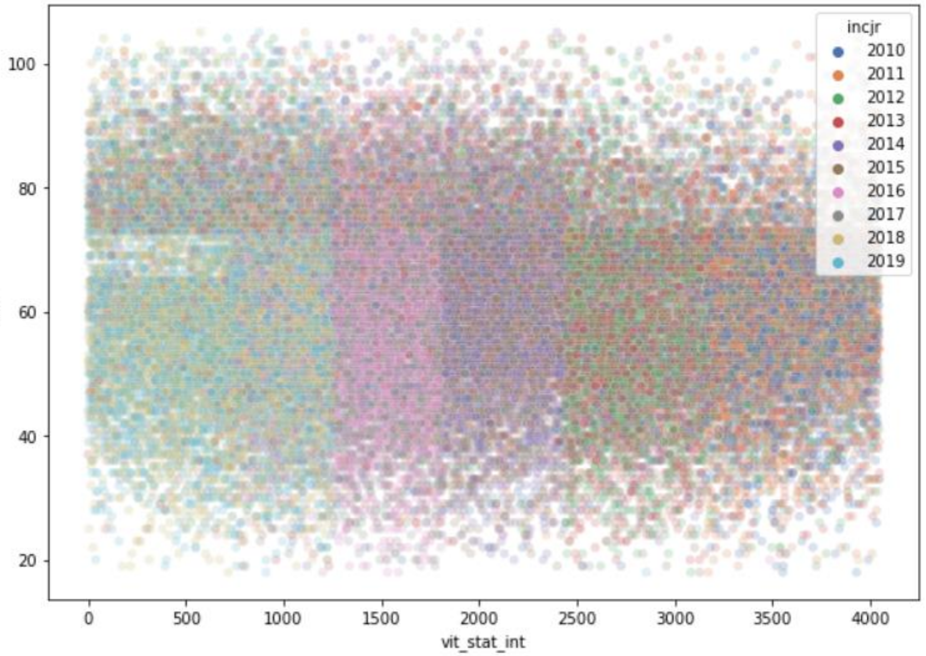
Time since incidence (vit_stat_int)

- One can notice that distribution of data points is not entirely random in neither of the sets. There are some patterns visible. There stands out a horizontal division at the value of around 73 at x axis, above which points with vit_stat of 0 prevail
- Vertical and horizontal „belts" are also visible
- Assumption is that these patterns are a side effect of the synthetic data generation process. It is acceptable if does not negatively impact correlations between features.

Original vs revised data set – overall view

hue = year of incidence

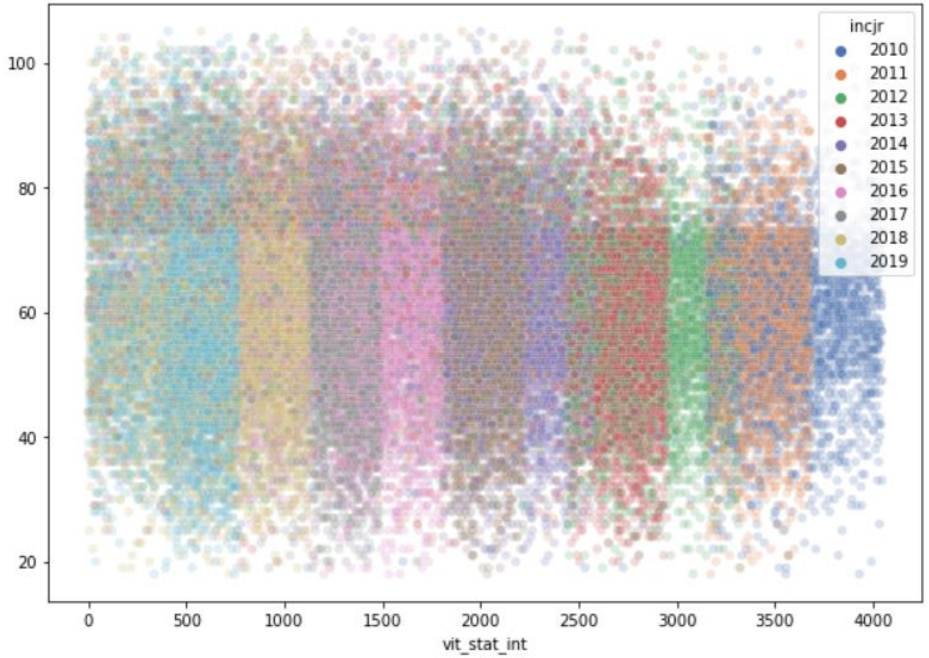
Original set



Age (leeft)

Time since incidence (vit_stat_int)

Revised set



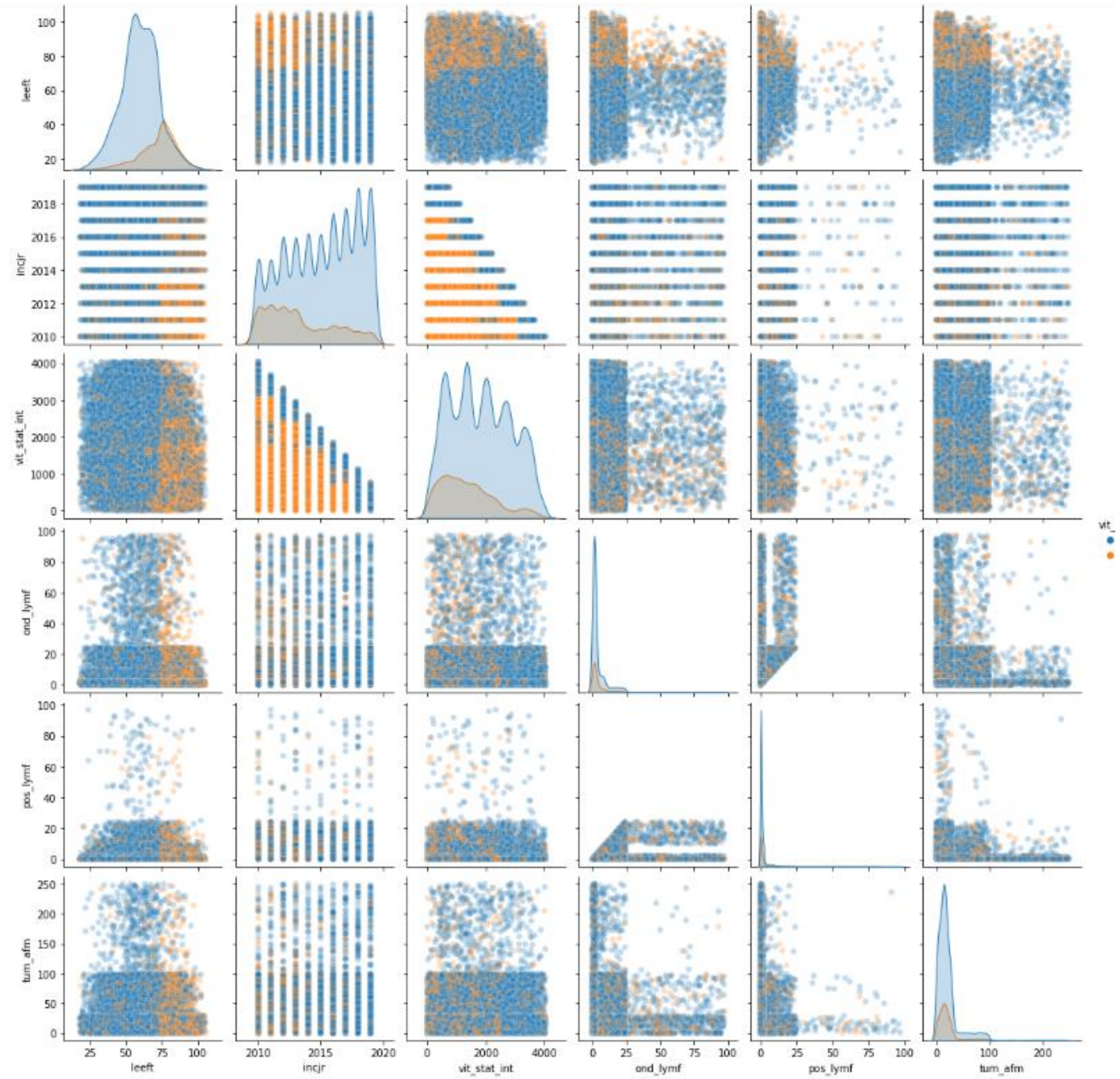
Age (leeft)

Time since incidence (vit_stat_int)

Similarly, some patterns are visible with regard to the density of data points above line of x= 73.

The vertical patterns can be explained by the way in which the dataset has been adjusted.

Revised data set - pairplot



The pairplot shows correlations between truly numeric features of which there are only a few:

- patient’s age
- year of incidence
- time (days) since incidence
- number of analysed lymph nodes
- number of positive lymph nodes
- tumor dimension (mm)

As can be seen, there are more patterns which give „artificial” impression.

One cannot determine a priori if this is an effect of the synthetic data generation, or some underlying rules of medical data coding play a role.

Cancer data analysis – KNR vs SEER

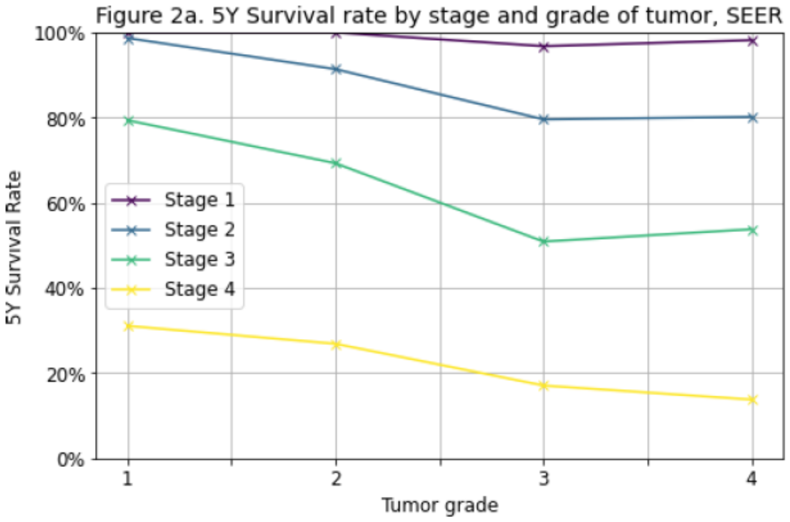
KNR vs SEER data

The comparison between NKR and SEER data (see previous report), has now been repeated with the revised dataset .

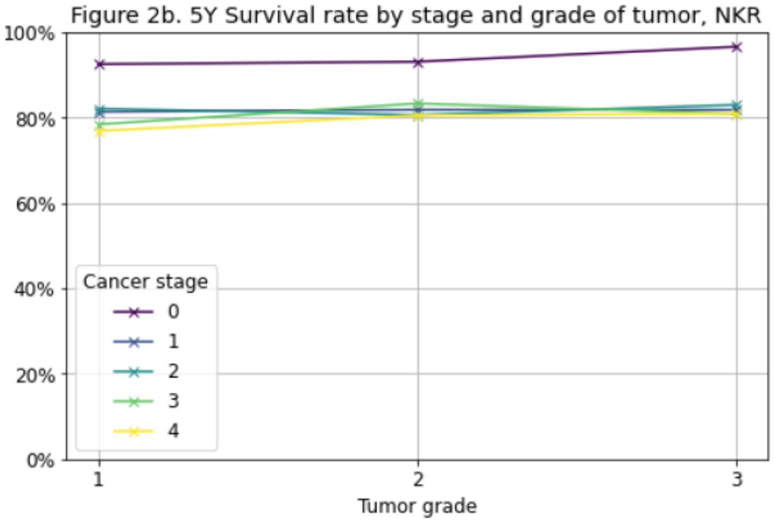
- previous analysis (March 2022): original dataset, 2010 cohort only (5241 records)
- present analysis: revised dataset, all the records for which 5 year survival rate could be calculated (32 460 records)

Observations:

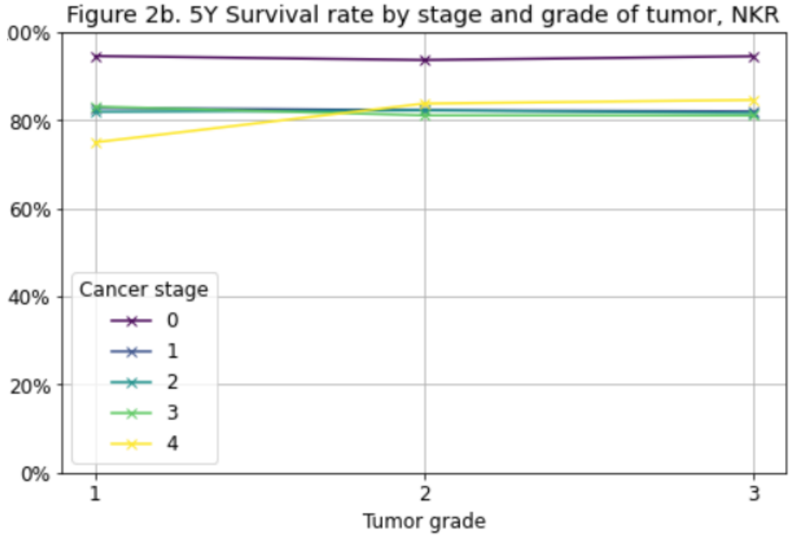
- very similar character of correlations
- some correlations are smoother (less scatter), most likely thanks to larger amount of data



SEER



KNR original set

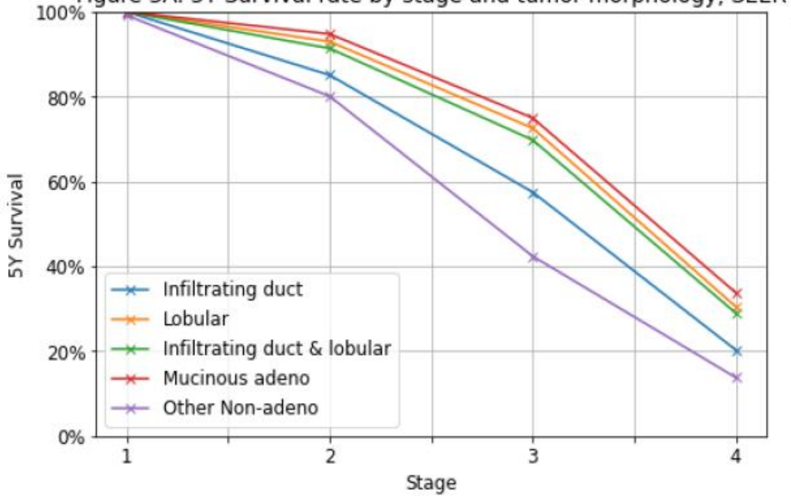


KNR revised set

KNR vs SEER data

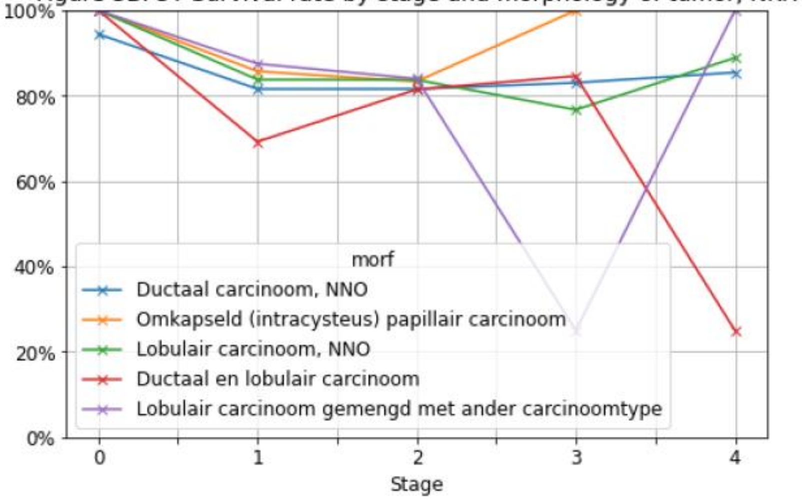
SEER

Figure 3A. 5Y Survival rate by stage and tumor morphology, SEER



KNR original set

Figure 3B. 5Y Survival rate by stage and morphology of tumor, NKR



KNR revised set

Figure 3B. 5Y Survival rate by stage and morphology of tumor, NKR

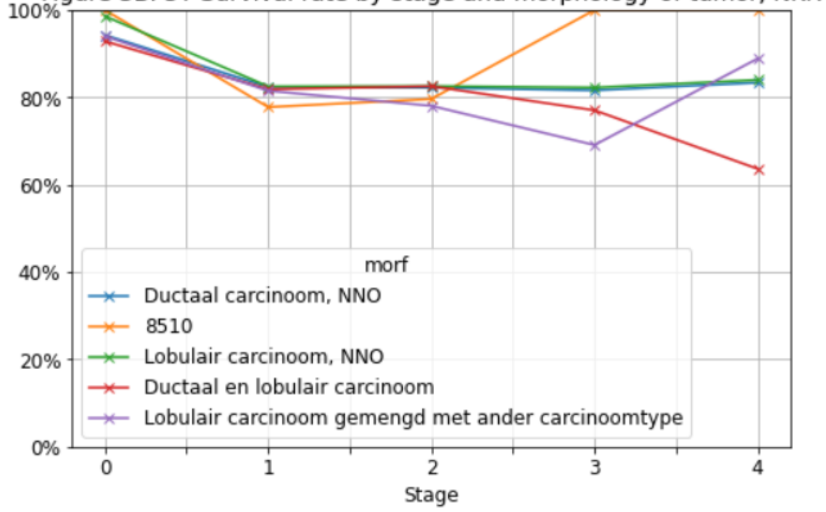


Figure 6a. Existence of Positive Nodes by Tumor Size, SEER

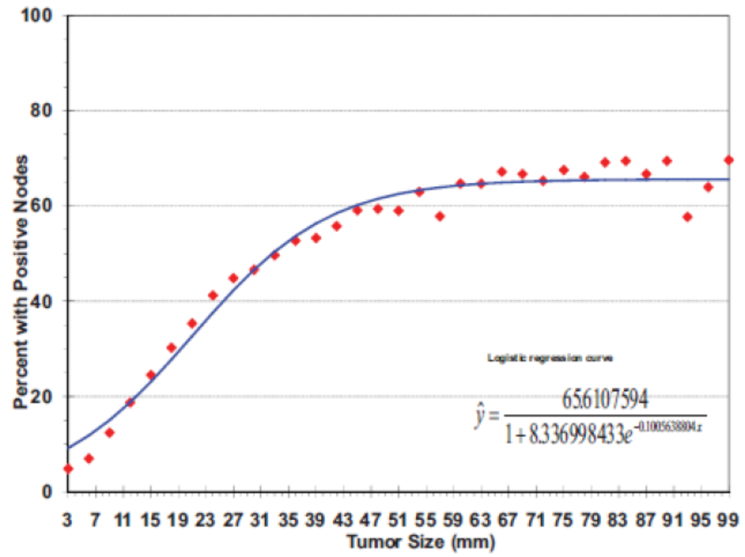


Figure 6b. Existence of Positive Nodes by Tumor Size, NKR

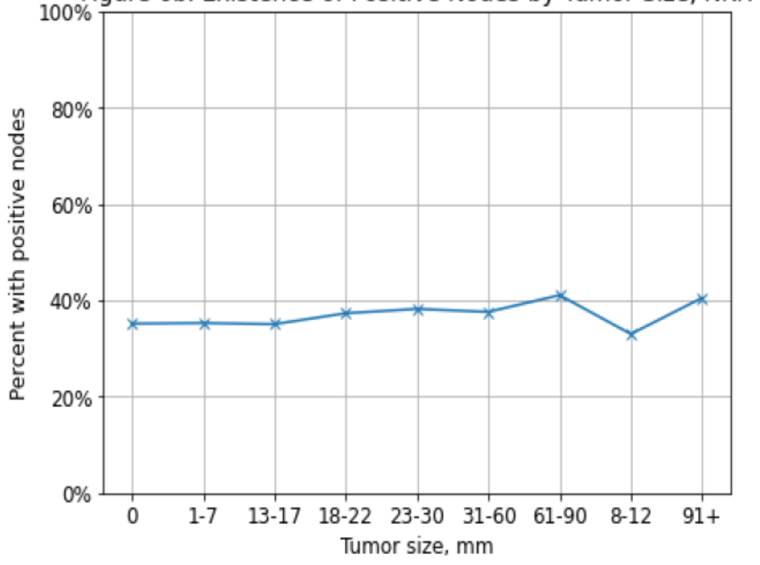
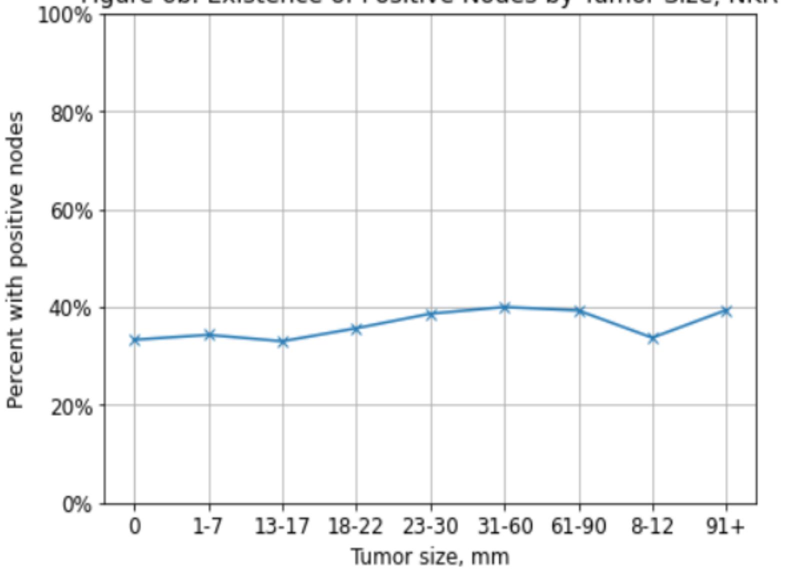


Figure 6b. Existence of Positive Nodes by Tumor Size, NKR



KNR vs SEER data

Conclusions:

- Differences between correlations derived from the original KNR set and the revised one are quite small. In some cases the relationships from the revised KNR set are smoothened, which may be due to bigger volume of data
- However, one can see again that the character of the relationships observed in SEER data is not visible in KNR data. For example, one can reasonably expect that 5Y survival rate would be lower the higher stadium of the cancer. Such relationship is clearly seen in SEER data, but hardly so in KNR data.
- There are two possible explanations for this:
    - Either in KNR data such relationships are not present
    - Or: Synthetic data do not adequately reflect the respective relationships in the original data
- One solution to clarify the underlying cause would be to repeat the analysis with the original KNR data

Modeling

Modeling

Goal - investigate the potential of ML models to predict patient’s outcome based KNR data

Two different modeling approaches were tested:

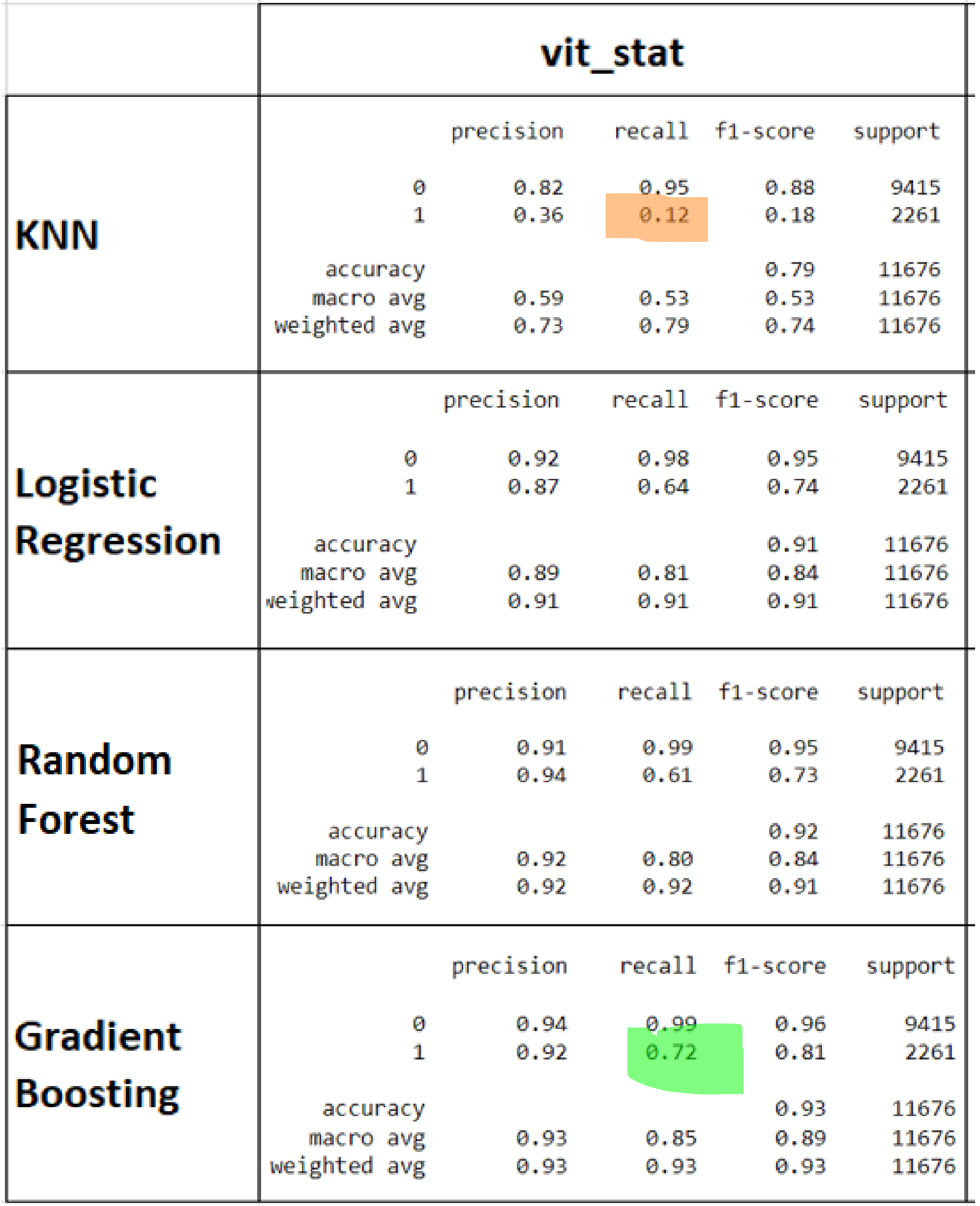
1. Target variable Y is **vital status of the patient (vit_stat).** Input features (X) are all features, including year of incidence (incjr), time since incidence (vit_stat_int), as well as patient’s age, medical conditions and treatment applied.
2. Target variable Y is **patient’s 5 year survival rate.** Input features (X) are patient’s data (incjr and vit_stat_int excluded).

Some more elaborate data preparation was applied in order to maximize the utilization of features, for example some categorical features have been converted into ordinal ones.

5 years’ survival rate could only be calculated for data until 2016.

The models used were logistic regression, KNN, random forest and gradient boosting.

Modeling – results of vital status models



vit_stat

KNN

```
              precision    recall  f1-score   support

           0       0.82      0.95      0.88      9415
           1       0.36      0.12      0.18      2261

    accuracy                           0.79     11676
   macro avg       0.59      0.53      0.53     11676
weighted avg       0.73      0.79      0.74     11676
```

Logistic Regression

```
              precision    recall  f1-score   support

           0       0.92      0.98      0.95      9415
           1       0.87      0.64      0.74      2261

    accuracy                           0.91     11676
   macro avg       0.89      0.81      0.84     11676
weighted avg       0.91      0.91      0.91     11676
```

Random Forest

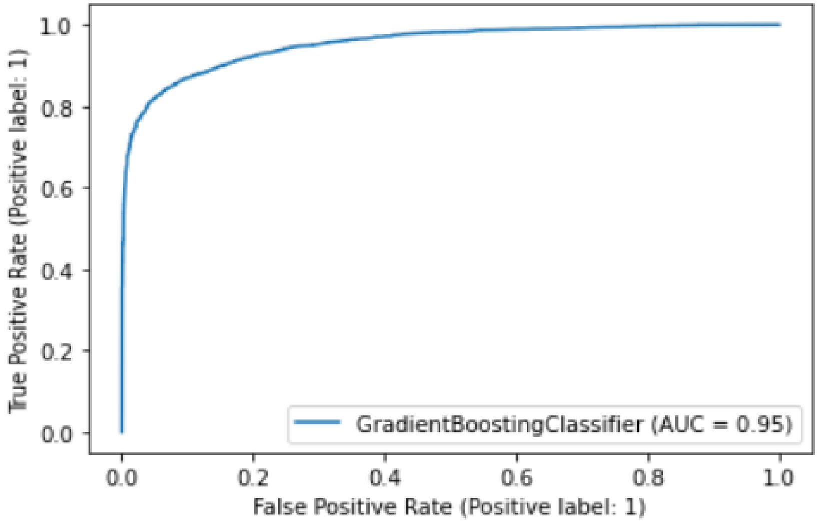
```
              precision    recall  f1-score   support

           0       0.91      0.99      0.95      9415
           1       0.94      0.61      0.73      2261

    accuracy                           0.92     11676
   macro avg       0.92      0.80      0.84     11676
weighted avg       0.92      0.92      0.91     11676
```

Gradient Boosting

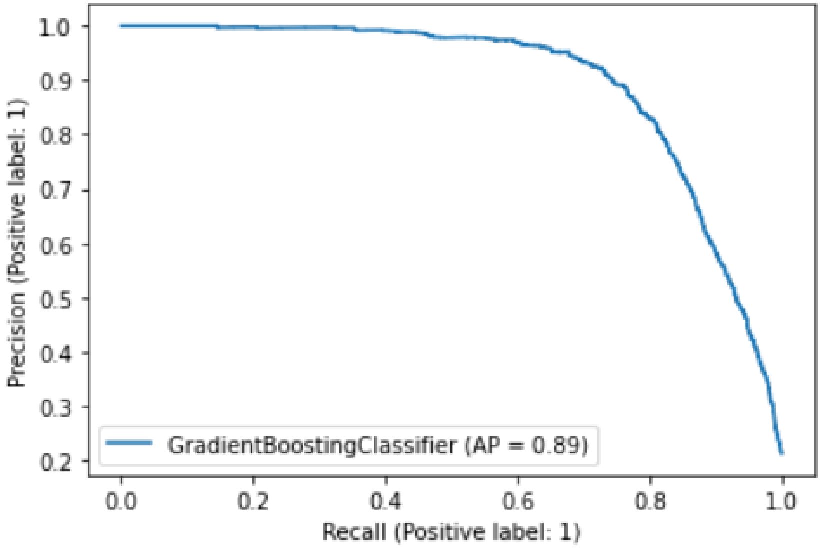
```
              precision    recall  f1-score   support

           0       0.94      0.99      0.96      9415
           1       0.92      0.72      0.81      2261

    accuracy                           0.93     11676
   macro avg       0.93      0.85      0.89     11676
weighted avg       0.93      0.93      0.93     11676
```

Gradient Boosting – best performing model

ROC curve

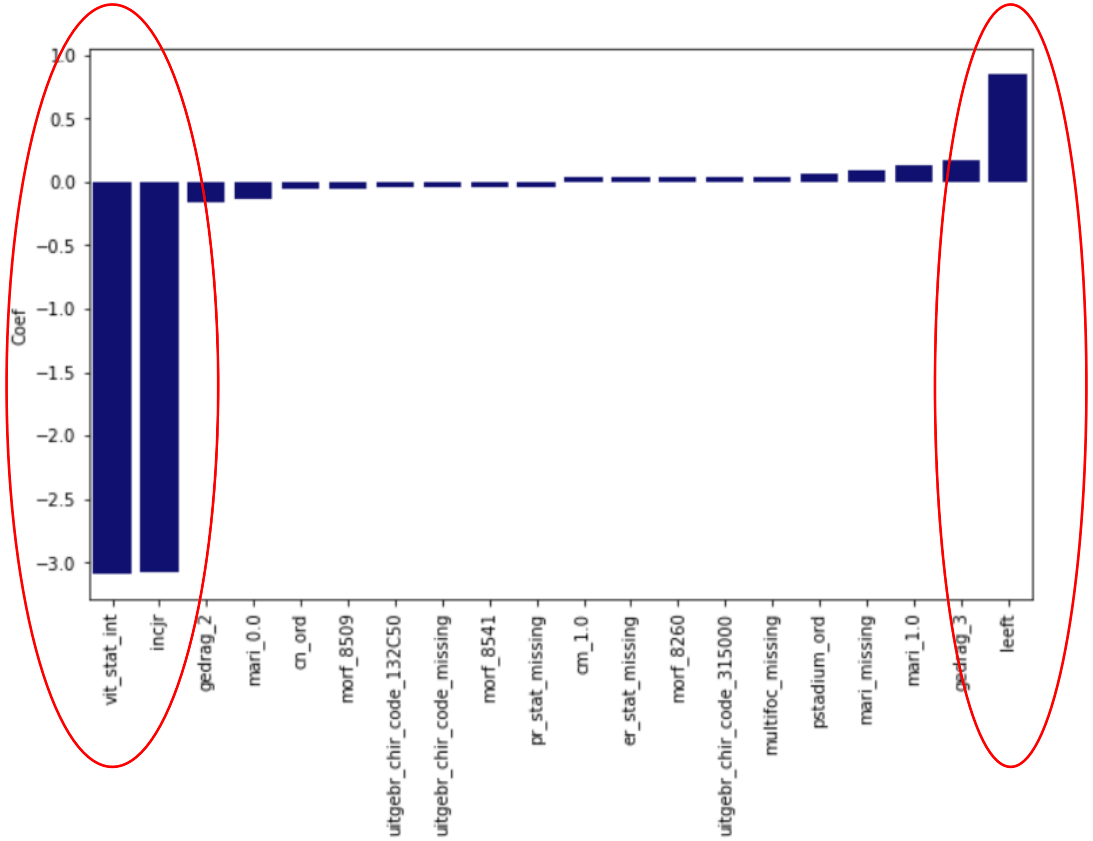


Precision - recall curve



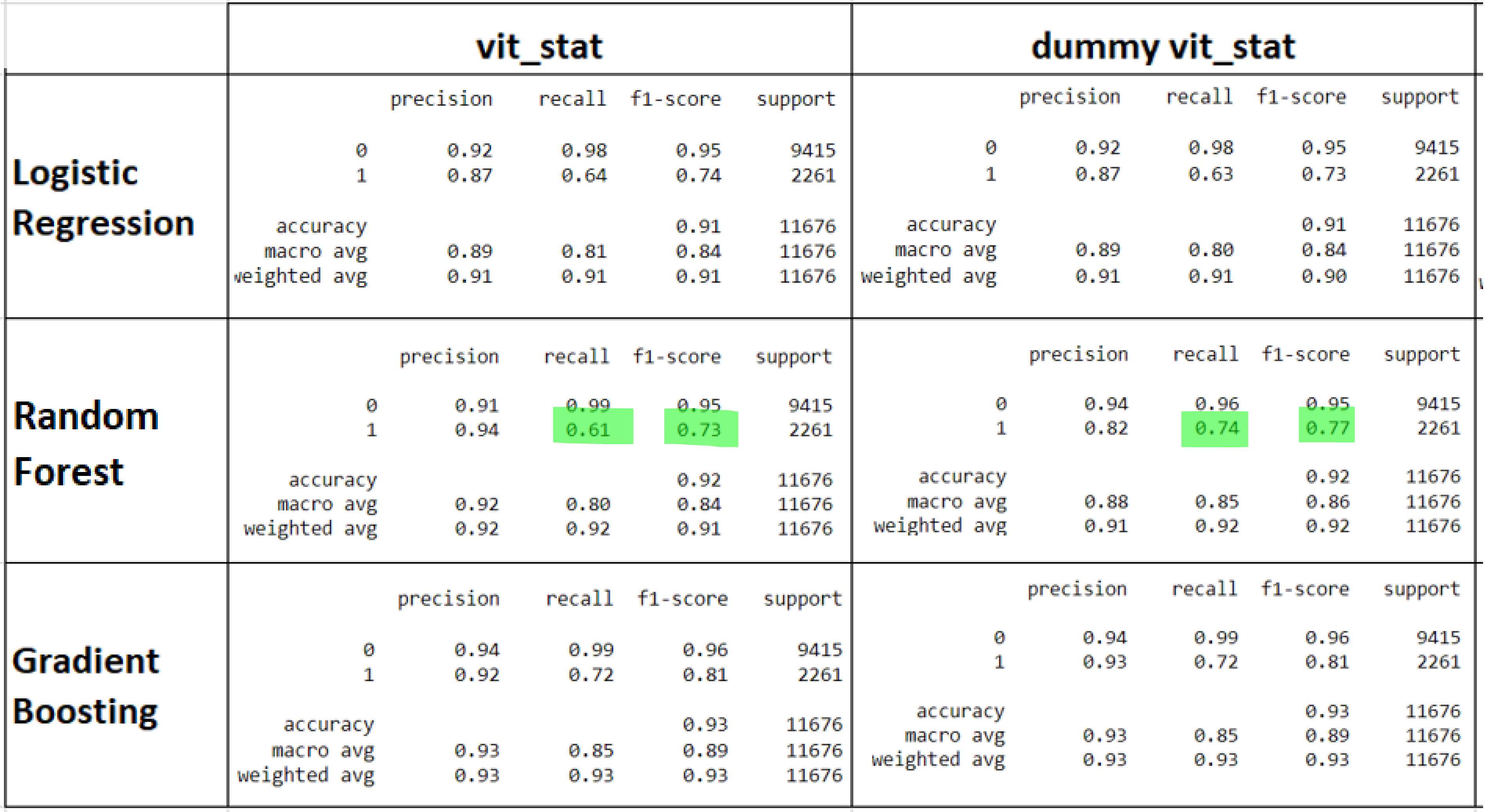
Modeling – results of vital status models - discussion

Linear regression coefficients



- Vital status models seem at first sight to perform quite well (except KNN)
- However, linear regression coefficients indicate **paramount impact of variables vit_stat_int (time since incidence), incjr (year of incidence) and patient's age**. It seems that clinical features (like cancer stadium and all others) play very little role. This is very unexpected and counter-intuitive
- To verify this observation, dummy models with only these 3 features have been tested

Modeling – vital status model performance - full vs „dummy” models



	vit_stat	dummy vit_stat

Logistic Regression

vit_stat

```
              precision    recall  f1-score   support

           0       0.92      0.98      0.95      9415
           1       0.87      0.64      0.74      2261

    accuracy                           0.91     11676
   macro avg       0.89      0.81      0.84     11676
weighted avg       0.91      0.91      0.91     11676
```

dummy vit_stat

```
              precision    recall  f1-score   support

           0       0.92      0.98      0.95      9415
           1       0.87      0.63      0.73      2261

    accuracy                           0.91     11676
   macro avg       0.89      0.80      0.84     11676
weighted avg       0.91      0.91      0.90     11676
```

Random Forest

vit_stat

```
              precision    recall  f1-score   support

           0       0.91      0.99      0.95      9415
           1       0.94      0.61      0.73      2261

    accuracy                           0.92     11676
   macro avg       0.92      0.80      0.84     11676
weighted avg       0.92      0.92      0.91     11676
```

dummy vit_stat

```
              precision    recall  f1-score   support

           0       0.94      0.96      0.95      9415
           1       0.82      0.74      0.77      2261

    accuracy                           0.92     11676
   macro avg       0.88      0.85      0.86     11676
weighted avg       0.91      0.92      0.92     11676
```

Gradient Boosting

vit_stat

```
              precision    recall  f1-score   support

           0       0.94      0.99      0.96      9415
           1       0.92      0.72      0.81      2261

    accuracy                           0.93     11676
   macro avg       0.93      0.85      0.89     11676
weighted avg       0.93      0.93      0.93     11676
```

dummy vit_stat

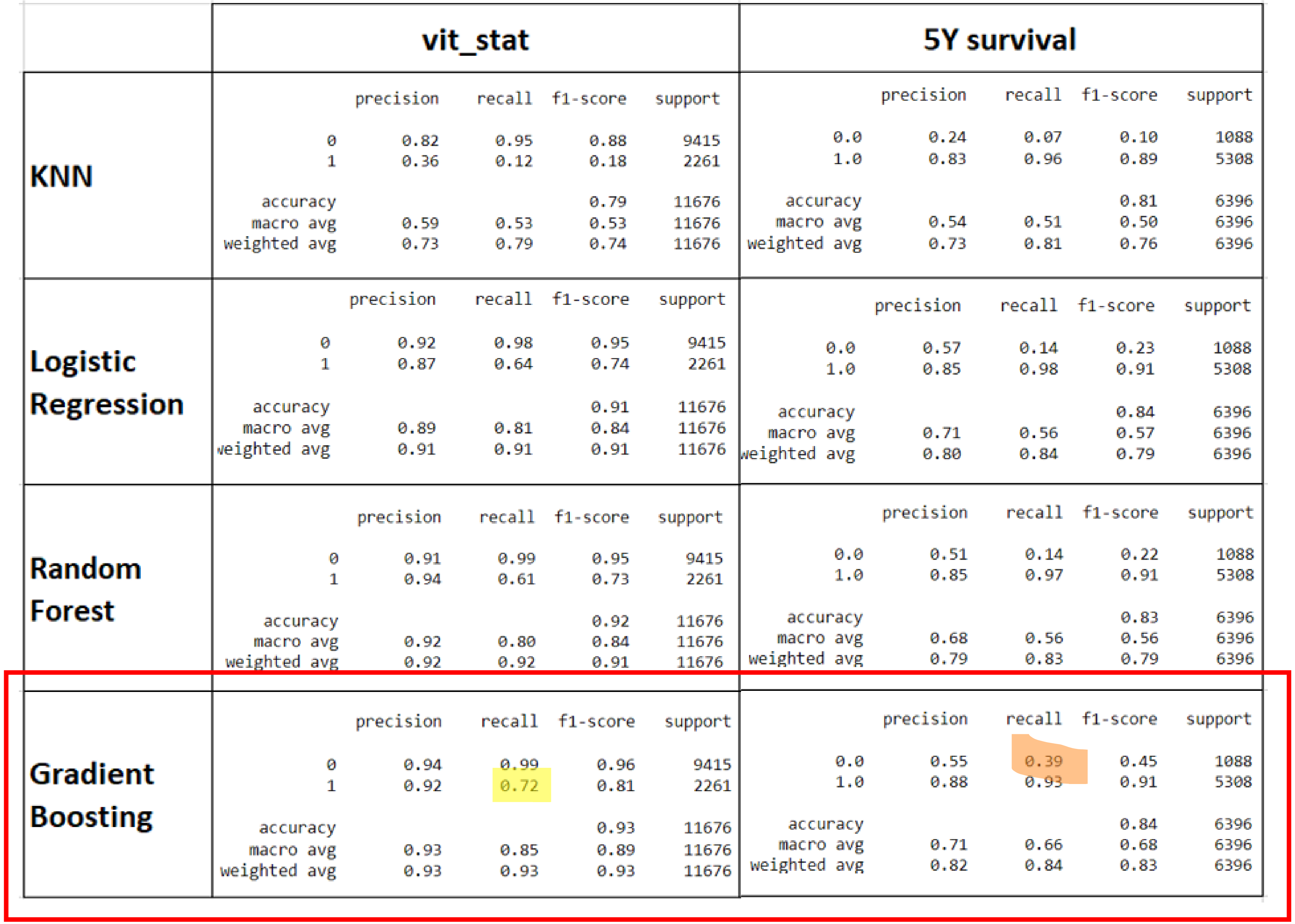
```
              precision    recall  f1-score   support

           0       0.94      0.99      0.96      9415
           1       0.93      0.72      0.81      2261

    accuracy                           0.93     11676
   macro avg       0.93      0.85      0.89     11676
weighted avg       0.93      0.93      0.93     11676
```

Conclusion: Performance of „dummy” models is the same, and sometimes better than complete models with all the features.

Vital status models vs Survival models

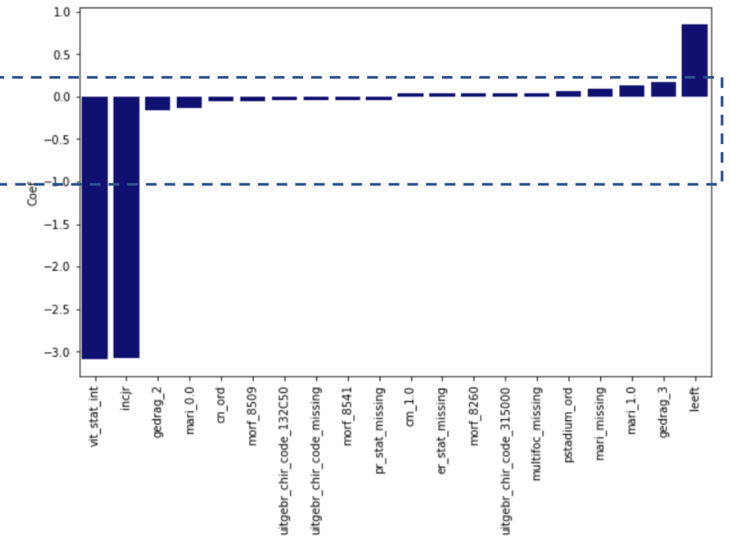
- **Vital status models** seem at first sight to perform quite well (except KNN)
- Precision and recall scores are much higher with respect to 0 value (patient alive) because the dataset is imbalanced (80 / 20 ratio of 0 values)
- Key issue: performance is driven by features which are not related to the patients medical condition. Model can predict the current vital status of a patient of age X (feature leeft), at whom a cancer has been diagnosed in year Y (incjr), which was Z days ago from now (vit_stat_int). However it says per se nothing about patient's chance of survival (or remaining period of life) after cancer detection
- The crucial impact of the incidence year is difficult to explain. One would expect that patient's vital status would depend on time passed since cancer incidence, but not on the year of incidence
- With these shortcomings of vital status models in mind, the goal has been set to build a **survival model**, e.g. a model predicting patient's survival after 5 years, a metric commonly used in cancer diagnostics.
- The respective target variable – survival_5Y, is constructed as patient's status at 1825 days (5 years) after incidence, and equals 1 if patient stays alive, or 0 if otherwise.
- Model types used were again logistic regression, KNN, random forest and gradient boosting.

Vital status models vs Survival models – Results

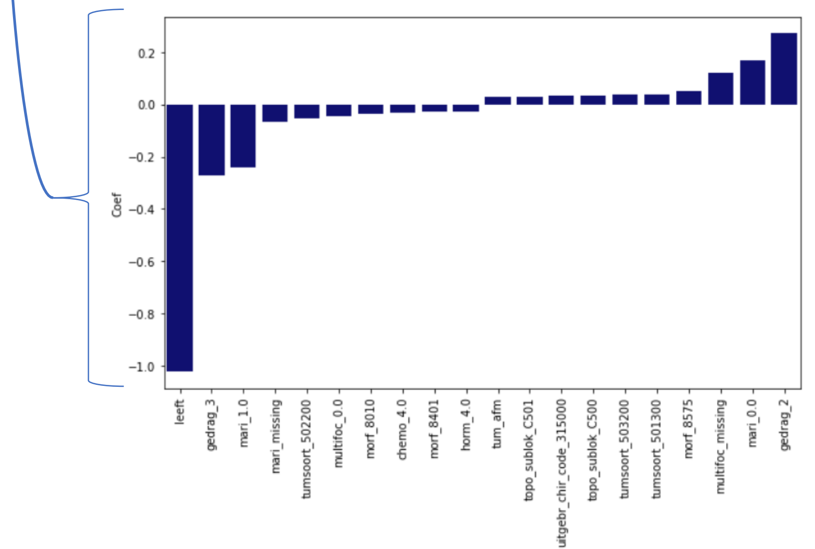


	vit_stat	5Y survival
KNN	precision recall f1-score support 0 0.82 0.95 0.88 9415 1 0.36 0.12 0.18 2261 accuracy 0.79 11676 macro avg 0.59 0.53 0.53 11676 weighted avg 0.73 0.79 0.74 11676	precision recall f1-score support 0.0 0.24 0.07 0.10 1088 1.0 0.83 0.96 0.89 5308 accuracy 0.81 6396 macro avg 0.54 0.51 0.50 6396 weighted avg 0.73 0.81 0.76 6396
Logistic Regression	precision recall f1-score support 0 0.92 0.98 0.95 9415 1 0.87 0.64 0.74 2261 accuracy 0.91 11676 macro avg 0.89 0.81 0.84 11676 weighted avg 0.91 0.91 0.91 11676	precision recall f1-score support 0.0 0.57 0.14 0.23 1088 1.0 0.85 0.98 0.91 5308 accuracy 0.84 6396 macro avg 0.71 0.56 0.57 6396 weighted avg 0.80 0.84 0.79 6396
Random Forest	precision recall f1-score support 0 0.91 0.99 0.95 9415 1 0.94 0.61 0.73 2261 accuracy 0.92 11676 macro avg 0.92 0.80 0.84 11676 weighted avg 0.92 0.92 0.91 11676	precision recall f1-score support 0.0 0.51 0.14 0.22 1088 1.0 0.85 0.97 0.91 5308 accuracy 0.83 6396 macro avg 0.68 0.56 0.56 6396 weighted avg 0.79 0.83 0.79 6396
Gradient Boosting	precision recall f1-score support 0 0.94 0.99 0.96 9415 1 0.92 0.72 0.81 2261 accuracy 0.93 11676 macro avg 0.93 0.85 0.89 11676 weighted avg 0.93 0.93 0.93 11676	precision recall f1-score support 0.0 0.55 0.39 0.45 1088 1.0 0.88 0.93 0.91 5308 accuracy 0.84 6396 macro avg 0.71 0.66 0.68 6396 weighted avg 0.82 0.84 0.83 6396

vit_stat model (LR)



5Y survival model (LR)



Survival models - conclusions

- **Metrics: Recall as preferred metric**
  - The most appropriate metric for survival model is Recall for 0 value (patient not survived 5 years). For the imbalanced dataset with ca 80% survived, this metric is the most selective.

- **Model performance – moderate.**
  - Recall for the survival models is in range from 0.07 (KNN) to 0.39 (gradient boosting). It is much lower than for vital status models (0.61 to 0.72). It can probably be somewhat improved by fine-tuning the model, but there seems to be a true limit in the data itself.

- **Impact of explanatory variables – under expectations**
  - Patient's age has the highest absolute linear regression coefficient (1.02). It is followed by medical features such as cancer bahavior (in situ or malignant, 0.27) and the mari procedure (0.24). Other medical features have much lower coefficients, of around 0.05 and less. This means that their correlations with patient's survival are very weak, even in cases when an obvious relationship should be expected, such as cancer stadium at detection (0.005 !)

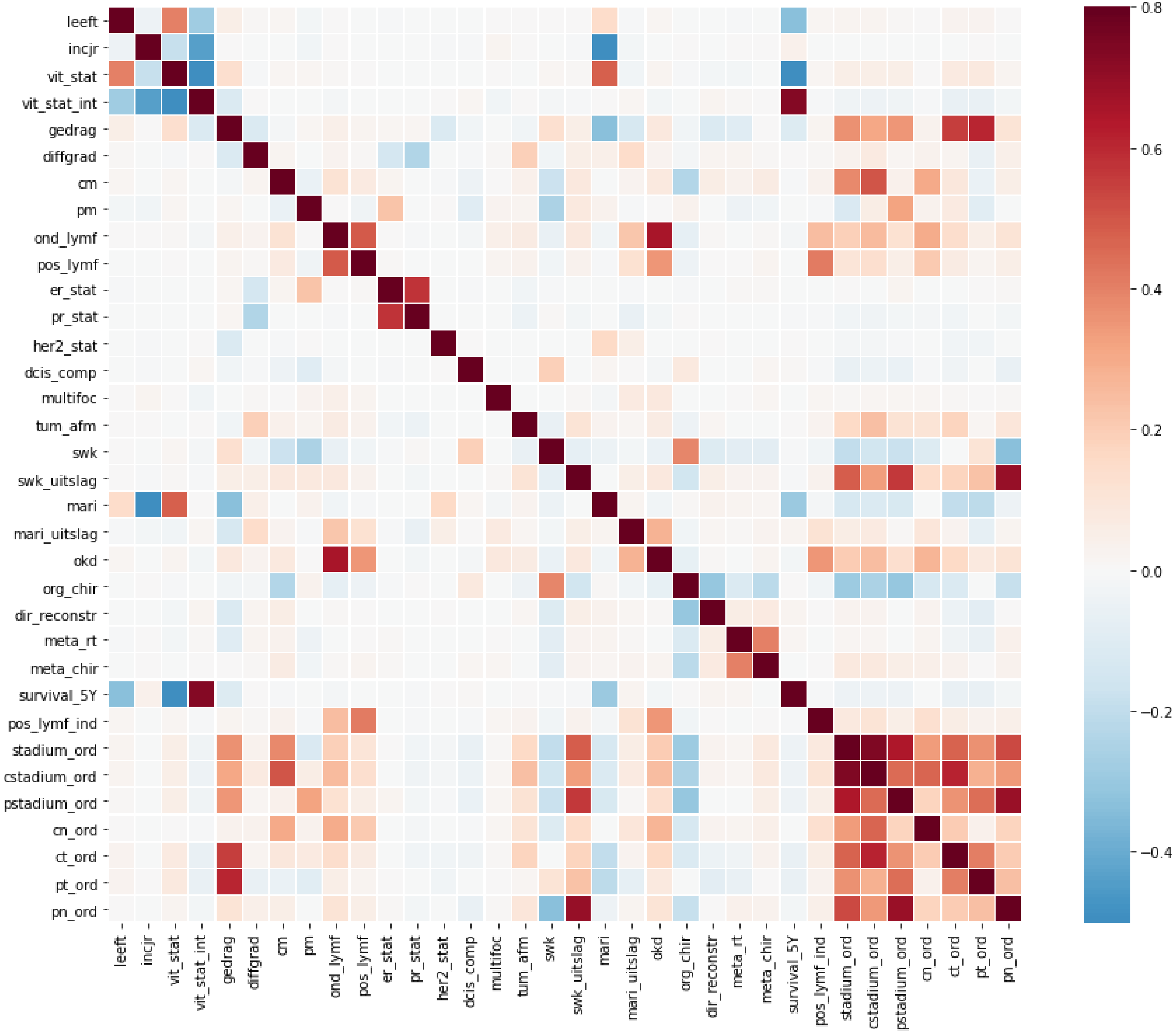
- **Reason of poor performance – hypotheses**
  - The challenge now is to explain why the majority of medical features have such a small influence on model predictions. For this, let us look back at correlations of the features with 5Y survival rate, as well as between the features themselves.

EDA revisited & Conclusions

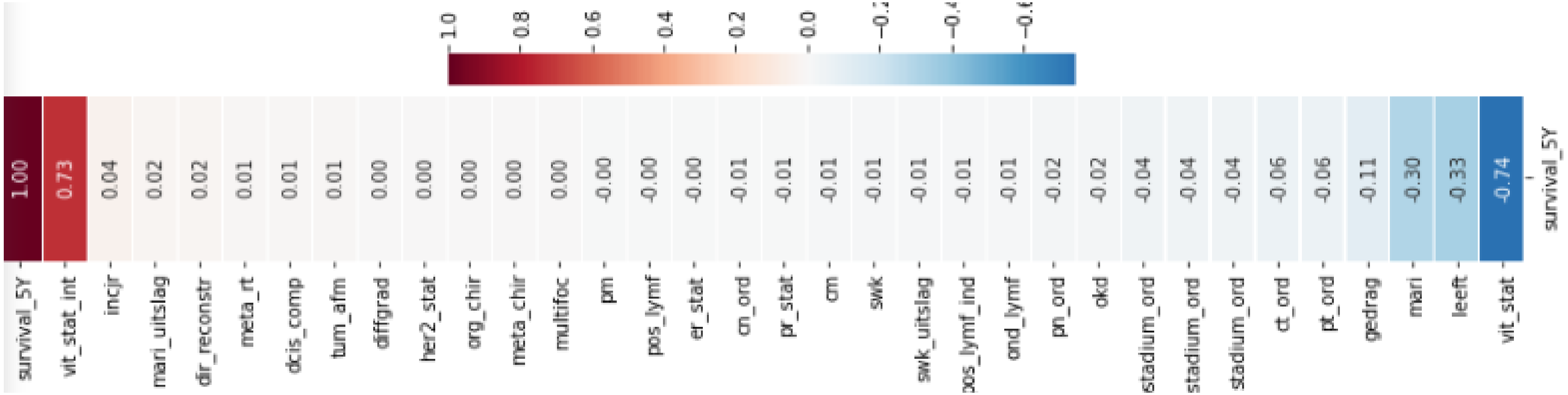
EDA revisited – correlation matrix

Conclusion: Confusing picture

- Corellations expected and clearly visible:
  - between various cancer stadium features
  - Between cancer character („gedrag”) and stadium
  - Between analysed and positive lymph nodes and stadium
- Corellations expected but missing or weak, for example:
  - Between differentation grade („diffgrad”) and stadium
  - Between er, pr and her2 stats and stadium
  - Between 5Y survival and most features (see next slide)
- Correlations not expected but present:
  - Between incidence year („incjr”) and mari procedure



5Y Survival rate – what does it depend on?

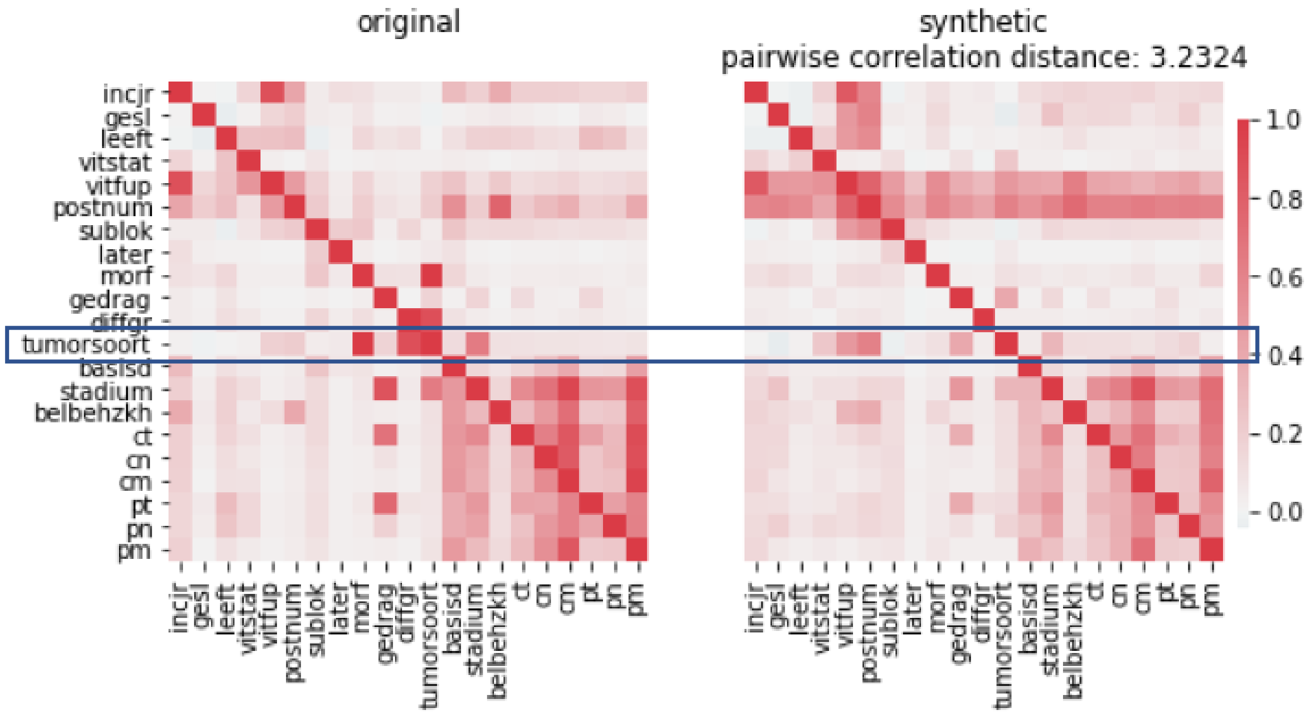


Apparently, the 5Y survival rate is only weakly correlated with clinical features:

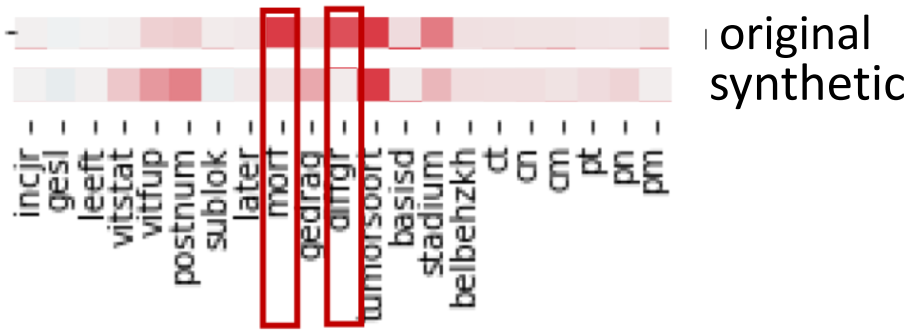
- mari procedure (0.30)
- cancer behavior (in situ vs malignant) (0.11)

Correlation with other clinical features is weak / non existent. This is the key issue to be explained.

Original vs Synthetic data – correlation matrix (report Daan)



Overall similarity, but ....



Strong correlations in the original, vs weak/none in the Synthetic:

- Tumor sort vs mophology
- Tumor sort vs differentiation grade

is anything lost in the synthetic data generation … ?

Way forward

- **Goal:**
  - Building model(s) with higher accuracy of 5Y survival prediction

- **Key challenge: explain missing correlations between clinical features and 5Y survival**
  - Resulting from synthetic data generation, or underlying issues with original data?

- **Possible next steps:**
  - Run EDA and models with the original dataset and compare results
  - Use SEER source data as reference