**CAS in Applied Data Science – University of Bern**

**Pathologically evaluated therapy induced tumor regression in non-small cell lung cancer**

**Guy Schnidrig, Philipp Zens**

**21st June 2021**

# Abstract

Lung cancer is one of the most common and most deadly cancer worldwide irrespective of sex. As projects for this CAS in applied data science, we planned a data analyses part investigating potential prognostic factors especially in neoadjuvant non-small cell lung cancer patients and machine learning part focused on the automated classification of tissue.

In this first data analysis part, we worked with 2 datasets. One is the publicly available TCGA dataset and the second one is a private dataset of neoadjuvantly treated non-small cell lung cancer patients at the Inselspital Bern. We were able to confirm the overall importance of the tumor stages for prognostic stratification of patients. These stages were important even after correction for additional prognostic markers as age, gender or smoking. Furthermore, we were able to show that the use of adjuvant therapy should be equal to neoadjuvant application. Finally, we were able to validate major pathological response, defined by the residual tumor in resections after neoadjuvant therapy, as an important prognostic marker which is currently in use as surrogate endpoint for clinical studies in resectable lung cancer patients.

Inhaltsverzeichnis

[Abstract 1](#_Toc75357703)

[1 Project Objectives 4](#_Toc75357704)

[2 Methods 4](#_Toc75357705)

[2.1 Software 4](#_Toc75357706)

[2.4 Analytical Methods 4](#_Toc75357707)

[2.4.1 Statistics 4](#_Toc75357708)

[2.4.2 Image Classification 5](#_Toc75357709)

[3 Data 5](#_Toc75357710)

[3.1 Datasets 5](#_Toc75357711)

[3.1.1 The Cancer Genome Atlas public dataset 5](#_Toc75357712)

[3.1.2 Bernese cohort of locally advanced NSCLC 5](#_Toc75357713)

[3.2 Digital Images 5](#_Toc75357714)

[4 Metadata 6](#_Toc75357715)

[4.1 TCGA data 6](#_Toc75357716)

[4.1.1 Web scraping information 6](#_Toc75357717)

[4.1.2 Variables Description 6](#_Toc75357718)

[4.2 Bernese Cohort 7](#_Toc75357719)

[4.2.1 Variables Description 7](#_Toc75357720)

[5 Data Quality 7](#_Toc75357721)

[6 Data Flow 7](#_Toc75357722)

[7 Data Model – only for TCGA data 8](#_Toc75357723)

[7.1 Conceptual Model 8](#_Toc75357724)

[7.2 Logical Model 8](#_Toc75357725)

[7.3 Physical Model 8](#_Toc75357726)

[8 Risks 8](#_Toc75357727)

[8.1 Information loss 8](#_Toc75357728)

[8.2 Interpretability of results 8](#_Toc75357729)

[9 Preliminary Studies 8](#_Toc75357730)

[9.1 AJCC tumor useful for the prediction of OS 8](#_Toc75357731)

[9.2 Neoadjuvant therapy has no survival benefit compared to adjuvant 9](#_Toc75357732)

[9.2.1 Age comparison 9](#_Toc75357733)

[9.2.2 Sex comparison 10](#_Toc75357734)

[9.2.3 Comparison of surgical procedure 10](#_Toc75357735)

[9.2.4 Differences 10](#_Toc75357736)

[9.2.5 Similar OS 11](#_Toc75357737)

[9.3 MPR is a prognostic marker 12](#_Toc75357738)

[9.3.1 Stage 12](#_Toc75357739)

[9.3.2 Surgical performance 13](#_Toc75357740)

[9.3.3 Demographic markers 13](#_Toc75357741)

[9.3.4 Independent prognostic significance of MPR 13](#_Toc75357742)

[10 Conclusions 14](#_Toc75357743)

# 1 Project Objectives

Lung cancer is one of the most frequent and lethal cancers worldwide. Recent years introduced new therapeutic options to metastasized lung cancer patients. However, in order to make these therapies promptly available for resectable lung cancer patients new endpoints are necessary for clinical trials. One such endpoint could be the assessment of residual tumor in resection specimen after the application of systemic treatments (Figure 1). This assessment can be rendered more objective and faster using digital image analyses tools based on neural networks. Thus, we had the following objectives for our projects during this CAS in Applied Data Sciences:

* Investigate prognostic biomarkers in non-small cell lung cancer (NSCLC) patients
  + Prognostic importance of the UICC TNM classification in early-stage NSCLC
  + Comparison of survival between patients with locally advanced NSCLC with or without neoadjuvant chemotherapy
  + Prognostic markers in neoadjuvant cases of NSCLC
* Training of a neural network for the identification of tissue types in Hematoxylin & Eosin (HE) – stained slides

# 2 Methods

## 2.1 Software

For the projects conducted during the CAS, a *GitHub* repository is available (<https://github.com/GuySchnidrig/CAS-2020-21>). This repository contains all necessary documents for the projects. However, due to privacy and because it’s considered to be sensitive data neither the clinical information nor the whole slide images and extracted patched of the digitized HE images are made available.

We used the *conda* package management system version 4.9.2 and *python* 3.8.5 for the data analytic project of module 1 & 2. All additionally required packages can be found a *requirements.txt* file in the repository.

For the image analyses, the open-source software *QuPath* version 0.2.3 was used for manual annotation of representative tissues. In order to train the network a *groovy* script was adapted to extract patches of 300 x 300 px from each region. The file names included the **image ID**, **class** and **coordinates of the patch** (Figure 2).

The *jupyter* notebook for tissue classification is supposed to be run using *Colab* for GPU supported training of the neural network.

## 2.4 Analytical Methods

### 2.4.1 Statistics

For the descriptive statistics mean with 95% confidence interval were used for normally distributed and median with interquartile ranges for non-normally distributed variables. Categorical variables were represented in crosstabs. The distribution of naturally ordered categorical variables and non-normally distributed continuous variables were compared using the non-parametric Mann-Whitney-U test. The distributions of normally distributed variables were compared using the Student’s t test. Correlation was assessed using the non-parametric Spearman correlation. Survival curves were depicted using Kaplan Meier plots and univariable analyses was conducted using the logrank test. For multivariable survival regression, we used cox proportional hazard models based on the proportional hazard assumption. Variables correlating with the variable of interest were not included in the multivariable analyses due to collinear effects. A p ≤ 0.05 was considered to be statistically significant.

### 2.4.2 Image Classification

We train a convolutional neural network for the correct classification of image patches. This is based on supervised learning, where in the training phase the correct label of the patch is given. In a first approach we will train a network from scratch and in a second approach we will finetune an existing network using our training data. The results of this part of the project will only be presented in the notebook of Module 3.

# 3 Data

## 3.1 Datasets

For the data analysis part, we will be working with two datasets: patient data assembled during *The Cancer Genome Atlas* (TCGA) program and a Bernese cohort of locally advanced NSCLC.

### 3.1.1 The Cancer Genome Atlas public dataset

The TCGA program was originally funded to better understand the molecular foundation of cancer. This program included sequencing, clinical and pathological data from different cancer entities. The clinical datafiles can be obtained manually and are structure in 3 *tab* separated files: *clinical.tsv, family\_history.tsv, exposure.tsv.* After merging, only prognostically relevant variables available for most cases were obtained. The final dataset includes 999 cases. Figure 3 and Table 2 represent the distribution of prognostically relevant variables of the TCGA cohort.

### 3.1.2 Bernese cohort of locally advanced NSCLC

The Bernese cohort was initially assembled based on the pathological files. This population consists of a study cohort of patients with neoadjuvant therapy before the resection of the NSCLC and a stage and histology matched control cohort of primary resected NSCLC. Stage matching was accomplished by including only patients with infiltrated loco-regional lymph nodes which qualifies for either neoadjuvant or adjuvant systemic therapy in addition to the surgical intervention. After the initial assemblage, the list was completed by considering the clinical files, contacting external physicians and the cantonal cancer registry. The cohort includes 246 cases corresponding to 245 patients, one patient was diagnosed with two synchronous tumors (LUSC and LUAD). Figure 4 depicts the association of MPR with other clinical parameters.

## 3.2 Digital Images

Appropriate Hematoxylin & Eosin (HE)-stained sections of the blocks with the most residual tumor were digitized for 228 patients. All files were digitized with the same scanner, Pannoramic 250 Flash III from 3D HISTECH (Budapest, Hungary), at x20 magnification. Representative regions containing **epithelial tumor**, **fibrous stroma**, **lymphocyte aggregates** and **necrosis** are manually annotated for the training of the neural network.

# 4 Metadata

## 4.1 TCGA data

## 4.1.1 Web scraping information

We manually downloaded the data using the official access tool for TCGA data <https://portal.gdc.cancer.gov/repository> applying the following filters:

* Availability of vital status
* LUSC and LUAD only
* Primary tumor site lung and bronchus
* Clinical data only
* *tab* separated files only

The database was accessed on 22nd June 2021. The obtained datasets were matched based on the *case ID* and reduced to the prognostically important variables.

An automated web scraping is possible, and an exemplary python script can be obtained in the documentation of the *GDC* portal. However, the web scraping results in 12 datasets, 6 for LUSC and 6 for LUAD, and we were not able to sort and join these databases in order to obtain a dataframe containing similar sample size and information as the manually downloaded ones.

## 4.1.2 Variables Description



## 4.2 Bernese Cohort

## 4.2.1 Variables Description



# 5 Data Quality

All clinical data included in the TCGA datasets was validated by the submitter. Thus, the correctness of the uploaded data is confirmed by the original data submitter. However, the TCGA is the program of choice to access large amount of publicly available clinical and molecular data of cancer patients. It is routinely used in the field of cancer research for preliminary analyses and as a comparison dataset for own findings. Therefore, we consider the TCGA program and the available dataset of high quality.

The data for the Bernese cohort was gathered by Philipp Zens in 2016. He was involved in the initial assemblage considering only the pathological files and further validation using clinical files and the cantonal cancer registry. All pathological evaluation of the cases was performed by Philipp Zens together with a trained pathologist specialized in lung pathology.

# 6 Data Flow

# 7 Data Model – only for TCGA data

## 7.1 Conceptual Model

## 7.2 Logical Model

## 7.3 Physical Model

# 8 Risks

The two major risks to consider are the one of **information loss** and **wrong interpretation of the results**.

## 8.1 Information loss

In order to be able to return to the original data, the two datasets used (TCGA, Bern) will not be changed during the analyses process. All additionally calculated variables or condensed categories will result in a new data frame within the *jupyter* notebook. The raw dataset of the Bern cohort is saved on the private computer of Philipp Zens and the raw TCGA tab separated datasheets are saved in their own folder in the GitHub repository.

In order to be able to previous versions of the code and for better collaboration during the project, all additionally created documents and analyses notebook are saved in the GitHub repository. Each of the collaboration will create an own branch and commit his changes in regular fashion. The branches will be merged to the main branch after regular periods.

## 8.2 Interpretability of results

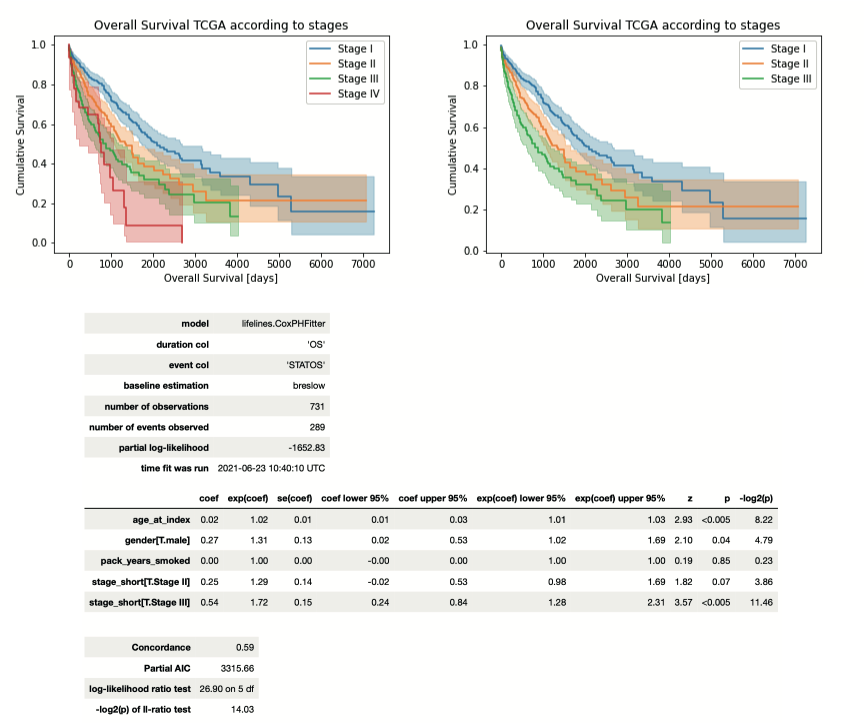
Both datasets are rather small compared to usual machine learning datasets. Even though, the quality of data is high for both sets, this will affect the validity of the regression models. However, most of the findings we describe in this project where already described in the literature and the analyses we perform can be considered for validation of already published data. Furthermore, we have only few samples but consider a broad range of potential confounders in the survival analyses and, thus, we can consider that a statistically significant regression truly indicates a prognostic marker (the p-value would rather drop when the analyses are repeated with larger sample size.

# 9 Preliminary Studies

## 9.1 AJCC tumor useful for the prediction of OS

The TCGA cohort has a median OS of 1528 das (95% CI 1379 – 1790 days). We further reduced the AJCC tumor stages to the four main stages (I, II, III, IV) in order to increase the sample size in each individual group. The AJCC tumor stages were able to stratify OS in the univariable analyses using the logrank test at p < 0.001. After consideration of the case numbers per stage and performing a pairwise logrank, we excluded the stage IV cases.

We further performed a multivariable analyses and could conclude that, besides higher age and male gender, the AJCC tumor stages remain statistically significant predictors of overall survival.



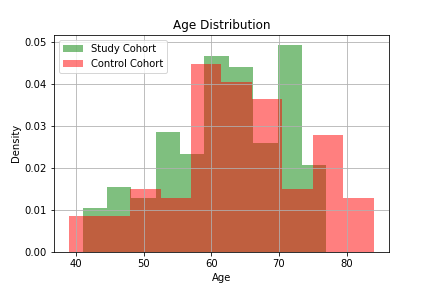
## 9.2 Neoadjuvant therapy has no survival benefit compared to adjuvant

In the Bernese population, the control cohort was matched to the study cohort in the sense that these patients would qualify for the neoadjuvant approach. There are still no clear guidelines when to use neoadjuvant or adjuvant therapy but multiple arguments are *pro* use of neoadjuvant regimens. We wanted to confirm the equality of OS in patients receiving neoadjuvant or adjuvant treatment.

Because the cohort was constructed retrospectively, we first needed to check if potential confounders such as age, smoking status or gender are similarly distributed in the two cohorts of interest.

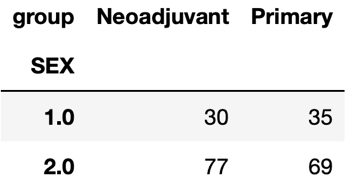
### 9.2.1 Age comparison

Mean age was 62 respectively 64 for the study and control cohort. The Student’s t test confirmed the similar distribution of this variable in both categories (p = 0.2).



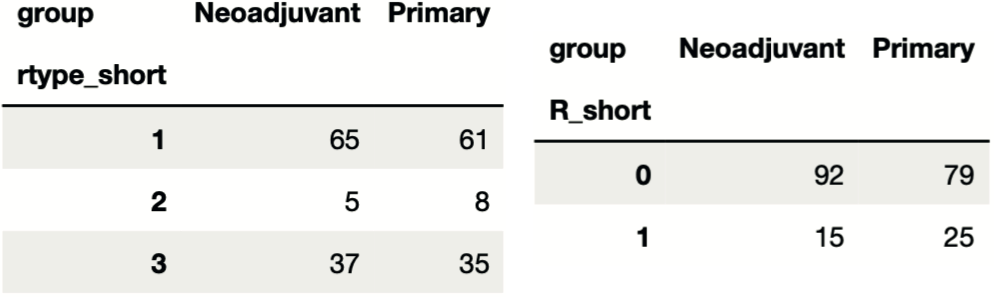
### 9.2.2 Sex comparison

In both cohorts, the majority of patients was male. The Χ2 test implied no difference in gender distribution (p = 0.463).



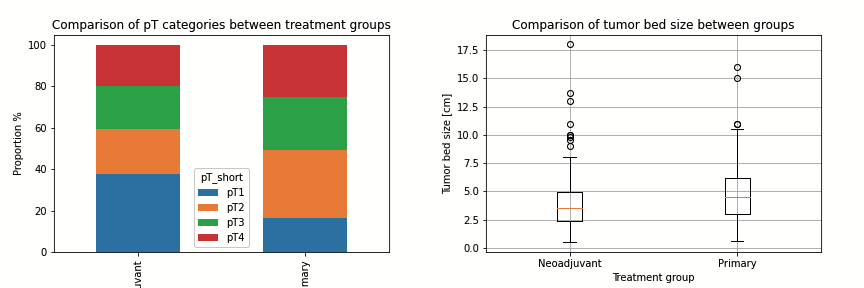
### 9.2.3 Comparison of surgical procedure

We further compared if the surgical performance, represented by the extent of resection (*rtype\_short*) and completeness of malignant tissue removal (*R\_short*), between the cohorts. Neither the performed surgical resection (p = 0.660) nor the completeness of resection (p = 0.079) seemed to be differently distributed.



### 9.2.4 Differences

As the control cohort was restricted to cases of stage III or higher, we investigated if the tumor bed and pT-categories were different between the cohorts. Tumor beds of the control cohort seemed to be larger (p = 0.001) and hence also higher pT-categories (p = 0.008) were observed in the control cohort. However, this is logic because the neoadjuvant therapy will result in a reduction of tumor size which is not always seen by a lot of fibrous scar tissue where the original tumor was.



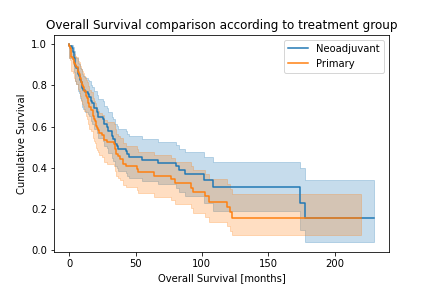
Furthermore, we see that significantly more patients received adjuvant chemotherapy in the control cohort (p < 0.001). We are happy, that this is the case because we wanted to compare survival between the two therapies strategy.

Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

### 9.2.5 Similar OS

Next, we looked at the survival curves of both cohorts. There is no significant difference in OS between patients receiving neoadjuvant or adjuvant chemotherapy (p = 0.298). Median survival in the study group (36.9, 28.2 – 82.7 months) was comparable with median survival in the control cohort (34.8, 20.1 – 51.4 months). Visually, however, the study cohort seemed to have superior OS especially considering the 5-year survival rate.

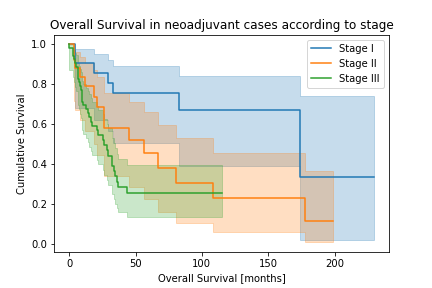


## 9.3 MPR is a prognostic marker

Next, we investigated additional prognostic markers in the study cohort, which could be potential surrogate markers for neoadjuvant clinical trials.

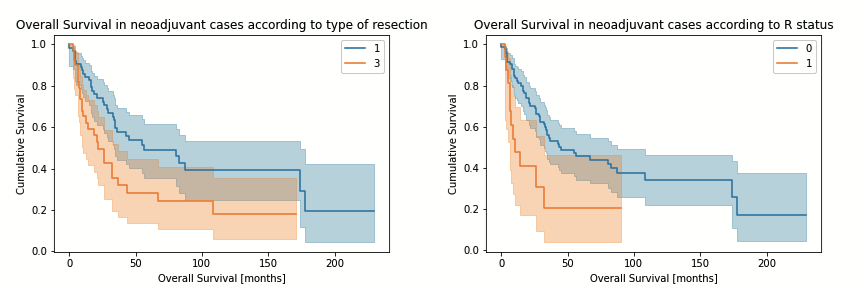
### 9.3.1 Stage

Tumor staged remained a prognostic marker as for primary resected patients (see TCGA analysis) with p = 0.009. We excluded Stage 0 patients because of limited sample size.



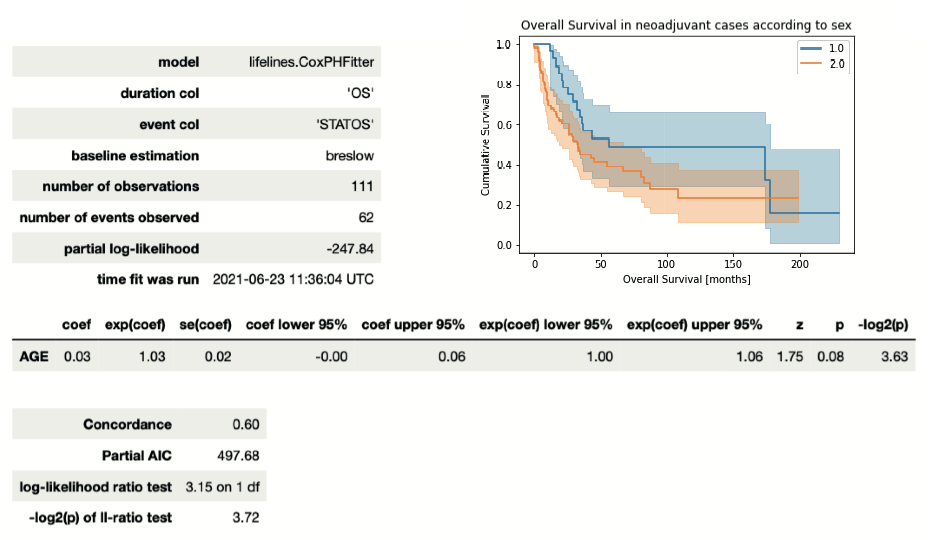
### 9.3.2 Surgical performance

The extent of surgery (p = 0.04) as well as the completeness of resection (p = 0.005) were prognostic in the neoadjuvant population.



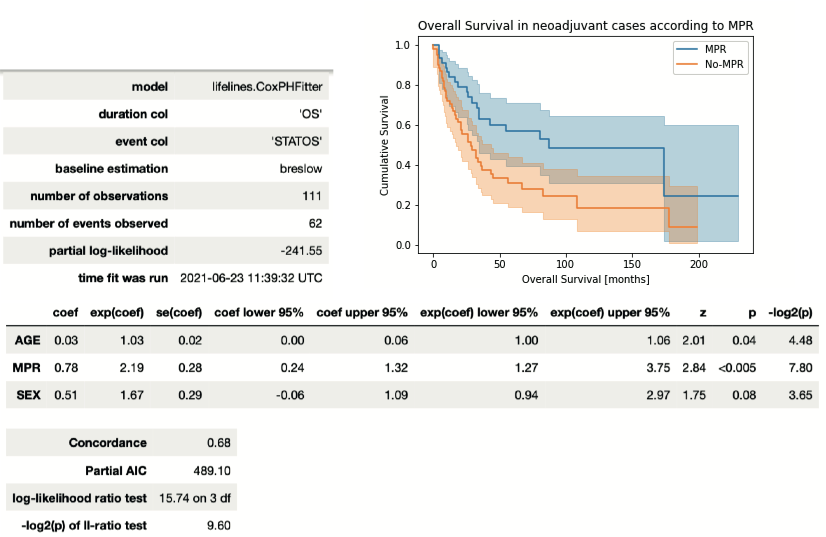
### 9.3.3 Demographic markers

Age (p = 0.08) and gender (p = 0.056) were just not significant prognostic markers. However, we included them in the multivariate model because they were markers in the TCGA analysis and our neoadjuvant cohort included much less patients.



### 9.3.4 Independent prognostic significance of MPR

MPR was a strong prognostic marker (p = 0.006) with patient showing MPR in the resection having a median OS of 87.6 months (34.3 – NA) compared to 28.2 months (17.2 – 43.4). Next, we applied again a multivariable model but did not include tumor stage, extent of resection or R-status because they correlated with MPR, hence, we wanted to avoid false results due to collinearity. In the multivariable model, patients without MPR had a twice as high hazard ratio of dying than people with MPR in the resection.



# 10 Conclusions

The pathologically determined tumor stages remain a very important prognostic tool in lung cancer patients. The stages are important as well for primary resected and neoadjuvantly treated patients.

It is irrelevant for the overall survival of the patients if systemic therapy is applied after or before surgery.

Major pathological response is a valid prognostic marker in neoadjuvant patients. It is currently used as surrogate endpoint in clinical trials for resectable lung cancer patients. However, there is a need of standardization for the assessment of this marker. Furthermore, a computer-based automatic estimation would be much more precise and faster than humans. Such an algorithm would be of tremendous value for routine diagnostic work.