**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. The development of new therapies targeting the hosts immune system rises new opportunities and new hope for patients suffering from this disease. In recent years, these advances have especially affected the metastasized lung cancer patients. But we are on the verge of introducing them to patient with surgically curable lung cancer.

Inhaltsverzeichnis

[Abstract 1](#_Toc75262909)

[1 Introduction & Objectives 3](#_Toc75262910)

[2 Methods 3](#_Toc75262911)

[2.1 Datasets 3](#_Toc75262912)

[2.1.1 The Cancer Genome Atlas public dataset 3](#_Toc75262913)

[2.1.2 Bernese cohort of locally advanced NSCLC 4](#_Toc75262914)

[2.2 Digital Images 4](#_Toc75262915)

[2.3 Software 5](#_Toc75262916)

# 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. 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 such as figures, jupyter notebooks etc. 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 the 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**.

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

## 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. 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.

### 2.4.2 Image Classification

We will 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.

# 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. This program includes also information of patients with NSCLC. The data can be accessed through a request script which defines different filters applied. Each

We retrieved the data through *python* based on a request scripts applying the following filter:

<https://api.gdc.cancer.gov/files> (original website for retrieval of the data)

* Available vital status (dead or alive)
* Histology (LUAD or LUSC)
* Primary tumor site lung
* Only clinical data
* *Tab* separated information

The search resulted in the retrieval of 12 files containing clinical data of patients with NSCLC. All retrieval requests are accompanied by a *MANIFEST.txt* containing the *id, filename, md5-hash, size and state* of the retrieved files.

### 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 necessitates 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). The following parameters were accessible:

|  |  |
| --- | --- |
| **Parameter** | **Format** |
| PID | Continuous |
| Age | Continuous |
| Sex | Categorical |
| Smoking | Categorical |
| Histology | Categorical |
| Bed | Continuous |
| Residual tumor | Continuous (1 – 100) |
| pT-Denominator | Categorical |
| pN-Denominator | Categorical |
| pM-Denominator | Categorical |
| UICC TNM stage | Categorical |
| Regression grade | Categorical |
| Group | Categorical |
| Neoadjuvant chemotherapy | Categorical |
| Number of neoadjuvant cycles | Continuous |
| Change of neoadjuvant chemotherapy | Categorical |
| Neoadjuvant radiotherapy | Categorical |
| Start date of neoadjuvant therapy | Date |
| Stop date of neoadjuvant therapy | Date |
| Date of resection | Date |
| Type of resection | Categorical |
| R-Status | Categorical |
| Adjuvant therapy | Categorical |
| State at last follow up | Binary |
| First Diagnosis | Date |
| Date of last follow up | Date |

The content of residual tumor in the initial tumor bed was evaluated by two investigators over the entire tumor bed and discordant cases were discussed on a bi-headed microscope. All cases were re-staged according to the current edition of the *UICC TNM classification of malignant tumors*.

## 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

# 5 Data Quality

# 6 Data Flow

# 7 Data Model

# 8 Risks

# 9 Preliminary Studies

# 10 Conclusions